Automatic classification of pathology reports

Automatic Classification of Pathology Reports

Automatic Classification of Pathology Reports

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Maastricht, op gezag van de Rector Magnificus, Prof. Mr. M.J. Cohen, volgens het besluit van het College der Decanen, in het openbaar te verdedigen op vrijdag 10 oktober 1997 om 14.00 uur

door

Lambertus Matthias de Bruijn

geboren op 4 december 1966 te Eindhoven

Promotor:
Prof. dr. ir. A. Hasman
Prof. dr. J.W. Arends
Prof. dr. ir. F.L. van Nes — *Technische Universiteit Eindhoven*

Beoordelingscommissie:
Prof. dr. ir. T. Arts (voorzitter)
Prof. dr. H.J. van den Herik
Dr. M. Nap — *Ziekenhuis De Wever en Gregorius Heerlen*
Prof. dr. R.W. Stockbrügger
Prof. dr. P.F. de Vries Robbé — *Katholieke Universiteit Nijmegen*

In de drukkosten van dit proefschrift werd bijgedragen door Stichting PALGA.

Dit proefschrift kwam tot stand in het kader van het instituut ExTra, deel uitmakend van de landelijke onderzoeksschool CaRe (Netherlands School of Primary Care Research), in 1995 erkend door de KNAW.
Het onderzoek dat beschreven wordt in dit proefschrift, werd uitgevoerd binnen het samenwerkingsverband tussen de Technische Universiteit Eindhoven en de Universiteit Maastricht.

AUTOMATIC CLASSIFICATION OF PATHOLOGY REPORTS

Contents:

'But is all this *true*?' said Brutha.

Didactylos shrugged. 'Could be. Could be. We are here and it is now. The way I see it is, after that, everything tends towards guesswork.'

'You mean you don't *know* it's true?' said Brutha.

'I *think* it might be,' said Didactylos. 'I could be wrong. Not being certain is what being a philosopher is all about.'

Terry Pratchett - Small Gods
(Corgi books, 1992)

CHAPTER · 1

INTRODUCTION

1 INTRODUCTION

1. General challenge and justification

The central question, that this thesis will attempt to present answers to, is: *'Is it possible to automatically assign classification codes to a diagnosis report that is written in natural language.'* This question will be explored in literature, and further answers will be sought in a new method for text classification. This method is presented, further investigated and evaluated in the various chapters of this work. Four aspects enframe the research that was performed: historical, practical, scientific and idealistic reasons.

1.1 Historical justification

In 1990, a collaborative study was proposed by the department of Medical Informatics at Maastricht University, and the department of Information Ergonomics of the Institute of Perception Research (IPO) at Eindhoven University of Technology. Under the title 'User Interfaces for medical systems', the question was posed whether automatic speech recognition and storage of digitized speech could solve the problem of data entry in medical information systems, and raise the quality of the stored information. The department of Pathology of the Academic Hospital Maastricht participated in the project. Their participation led to the adoption of pathology diagnostics as the medical working terrain for the project.

At IPO, Ellen Verheijen concentrated her research on user interface aspects of speech recognition. Her experiments revealed that pathologists have trouble detecting recognition errors in dictated texts, be it right after dictation or the day after dictation. She concluded that application of speech recognition is still inadvisable, because the error rate of dictation systems is still many times higher than for secretaries (5-10% vs. less than 1% for secretaries). On February 7 1997, Ellen Verheijen successfully defended her thesis (Verheijen, 1997).

The Medical Informatics part of the research was to concentrate on the informational side of diagnosis reporting. The state of the art on speech recognition technology was explored and reported in (De Bruijn et al. 1995, included as appendix 2 of this thesis). The speech technology market developed slower than was foreseen in 1990, so we could not obtain practical experience in this field. Instead, research was done on new informational functions that would be accessible through an information system. The initial sidetrack of assessing the possibility of automatic coding was found to contain such interesting aspects that it was decided to dedicate more research on this topic. The result is this thesis.

Practical justification 1.2

Pathology laboratories have an enormous throughput. The laboratory of the Academic Hospital in Maastricht has a yearly production that lies around 25,000 examinations. With a staff of 28 laboratory technicians, 72 secretaries and 26 scientific staff (who dedicate part of their time to research and teaching), it can be considered moderate sized for the Dutch situation. Appendix 1 lists some characteristics of the production of pathology, including a division by the kind of material that is processed and statistics of the diagnosis reports.

Encoding cases is one of the activities involved in diagnosing material and reporting about it. It is a burdensome task: a thesaurus must often be consulted which contains, in the version for Dutch pathology, about 15.000 words. Decisions must be made on which of the terms would match the case best. The task is susceptible to impatience or laziness (cf. Hall and Lemoine 1986).

Berman and Moore (1996) write: 'Anatomic pathologists are among the most prolific prose writers. On the basis of their output [4000 reports each year] anatomic pathologists far exceed the productivity of virtually all professional writers'. Any structural improvement per examination has an impact on the pathologist's daily task and a large impact on nationwide productivity.

Scientific justification 1.3

The method that is introduced in this work, is new to medical narrative analysis. The literature review (chapter 2) shows that other research has been reported on automatic classification of diagnosis reports, but these methods either depend on (extensive) domain modelling or on the language that is used in the reports. The new method has no dependencies on domain, language or classification system so it might as well be used to classify real estate taxation reports in Italian. This characteristic is kept intact as much as possible in experiments with the method in order to make transfer of the results as valuable as possible. Only in the last instance, additional procedures are tailor made on the domain so that weak points of the method are shored. Additional procedures enable implementation of the method in a usable setting.

Idealistic justification 1.4

Epidemiologic data is a potential source for clinical decision making. As Berman and Moore (1996) write: 'In many cases, simple demographics point so closely to the identity of a lesion that most pathologists would be reluctant to make a diagnosis when a case does not conform to an expected presentation. For instance, a pathologist might hesitate before making the diagnosis of Ewing's tumor in a black patient'. Data that is encoded in formal structures such as SNOMED or ICD can be the key to perform studies on a large scale, or identify occurrences of rare diseases in a large population. In The Netherlands, the PALGA database (PALGA stands for Pathologisch Anatomisch Landelijk Geautomatiseerd Archief - Dutch Network

and National Database for Pathology) which stores SNOMED based classifications, is used to feed the National Cancer Registry. PALGA's annual reports (1992, 1995) list a large number of nationwide queries on the database. Results of these queries were used in a number of published studies (cf. PALGA 1992).

The quality of the filed epidemiologic data depends on the quality of the database, which in turn depends on the quality of encoding the pathological findings. Errors are being made in (manual) data coding: this is a recognised fact that will also be addressed in chapters 2 and 5. Classifications should be complete, concise, correct and consistent. If automatic encoding or automatic support in manual coding could reduce error rates and improve coding quality, then a small contribution can be made to the overall quality of medical care.

2. Contents of this thesis

In this thesis, the possibility of automatic classification of pathology reports is explored. The work concentrates on three domains: (1) construction of a classification method and comparing a number of different design options; (2) exploring the difficulties in evaluation and proposing solutions to these difficulties; and (3) actual evaluation of the classification method

The rest of this introduction gives some clarification on what pathology is, notably from an informational point of view, and what processes are involved in making and reporting a diagnosis. The SNOMED classification system is described, and the use of this classification language in the PALGA database is discussed. These sections may be skipped by the medical professional.

Chapter 2 gives a literature overview of medical language processing and automatic classification of free narrative. Several studies are reviewed that addressed issues in automatic coding. The foundations in literature are discussed with regards to the approach in automatic classification that is used in the further experiments.

An extended description of the automatic classification system is given in chapter 3. It discusses the technical implementation, gives the range of options and defends parameter choices. This system was put to work: data was collected in a pilot experiment which consisted of an expert evaluation and initial simulations.

The pilot experiments led to expansion of the data collection: a larger sample from the Maastricht laboratory was obtained, and an additional collection was gathered from another laboratory. With these collections, a number of further simulations was performed. The results of these simulations are given and discussed in chapter 4.

Since the results of the simulation experiments were promising enough, it was decided that a large scale expert evaluation was justified. This experiment involved the reading of 240 reports by 18 subjects, and them rating a total of 1200

classification lines. This experiment is discussed in chapter 5. This chapter also describes how comparison of code strings can be used to predict expert ratings. Code string comparison can then be used to establish an artificial judgment measure for assessing further simulations.

Chapter 6, finally, discusses the total of the study. In this chapter, the simulations are realigned on the outcomes of chapter 5 and practical solutions for weaker points of the method are discussed and put to the test. The important issues of the user interface are elaborated upon. After a short enumeration of possible following experiments, the final conclusions of this research are given.

Definitions and clarification 3.

What is pathology 3.1

Clinical pathology is a diagnostic service: at the pathology department, tissue or cell samples are examined on order of an attending physician. The physicians who request laboratory examinations wish to confirm a clinical impression or establish a diagnosis, rule out a diagnosis, monitor therapy, establish a prognosis or screen for or detect disease (Woo and Henry 1991). Other diagnostic services that physicians have at their disposal include clinical laboratories (such as bacteriology and chemistry) and diagnostic imaging laboratories (including radiography and MRI) – see also table 1. Of the diagnostic resources, histopathological examination usually forms the *golden standard* because most disorders are reflected in specific structured alterations at the level of the cell tissue.

Patient care service is one of the tasks of a laboratory department, along with administration, teaching, and research (Woo and Henry 1991). Patient care service is the task that I concentrate on here, and it consists of the following phases, which may have an iterative character:

◊ indications and selection
◊ technology and generation (preparation, staining, imaging)
◊ interpretation and translation (reporting) (after Woo and Henry 1991).

The tasks lead to an informational view, in which Woo and Henry's patient care service is abstracted into a 'black box' 'into which one puts requests and specimens, and from which are produced reports from test results (and bills for those tests)' (Aller 1991). A more complete view shows that patient information and population data may be needed at the input side. The black box model assumes that consulting text books or conferring with collegues – which may be necessary to reach conclusions – takes place within the black box.

On the output side, several reports may be distinguished: the primary report goes to the physician that requested the examination; speed is a key factor here and it appears that the 'conclusions' line is the most important part. A secondary

destination is the patient record, which requires a complete and detailed description of observations, maybe even of those that did not contribute directly to the final conclusion. A tertiary destination is the scientific database that is used to support clinical and epidemiological studies. This requires a short and standardised summary of the findings so that a full report can be retrieved and which allows statistical processing of the abstracted data. The last two destinations are the department's own archive and the administration system. Figure 1 summarizes the information flow around a diagnostic department.

3.2 Opening the black box... pathology diagnostics

At the black box from figure 1, material and data comes in and reports are produced to be sent out. In between, processing is performed in two streams which may have an iterative character. There is a technical path that involves the handling and processing of the material so that observations can be made. Along that path lies a cognitive trajectum: hypothesis forming, information collection, perception (making observations) and hypothesis testing.

Pathology distinguishes

- ◊ histology examinations: tissue material (histos = woven tissue, logos = study)
- ◊ cytology examinations: cell material in a fluid environment (cytos = cell)
- ◊ autopsy: post-mortem examination.

For histology, material is cut and subjected to gross examination, i.e. macroscopically, with the bare eye – which may include measuring and weighing. Small parts are included as a whole, larger parts are dissected following certain procedures and

TABLE 1: diagnostic services; after Griffith (1987).

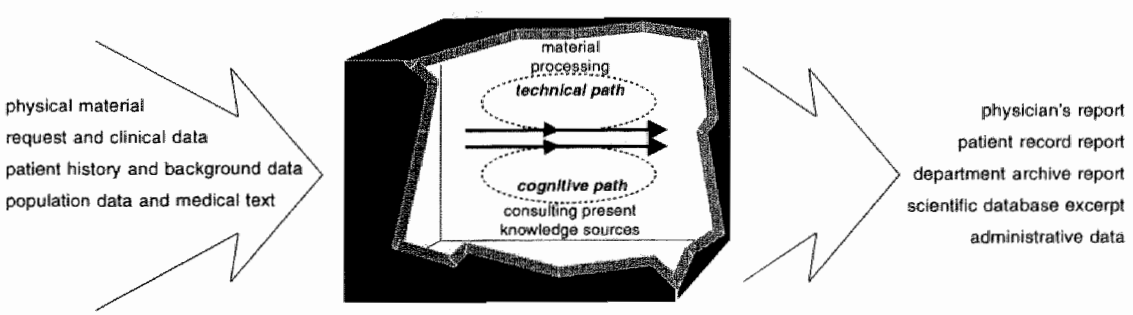
- ♦ clinical laboratory
  - ◊ chemistry
  - ◊ hematology (and coagulation)
  - ◊ (histo)pathology
  - ◊ microbiology/bacteriology, virology and immunology
  - ◊ autopsy and morgue
- ♦ diagnostic imaging
  - ◊ radiography
  - ◊ (computerized) tomography (CT)
  - ◊ radioisotope studies
  - ◊ nuclear / magnetic resonance imaging (MRI)
  - ◊ ultrasound / sonography
- ♦ cardiopulmonary laboratory
  - ◊ electrocardiology
  - ◊ pulmonary function
  - ◊ heart catherization
- ♦ other
  - ◊ electroencephalography
  - ◊ electromyography
  - ◊ audiology

guidelines depending on the kind of material. The small parts and the cut-outs of the large parts are processed into slides, which is done overnight. The next day, the slides are examined with a microscope and a diagnosis is formed. If it is not possible to ascertain a hypothesis, the use of additional markers or alternative staining may be requested.

Bloom and Fawcett (1975) describe the methodological background of histology: 'the ideal histological method [of examination] would be one that resulted in minimum deviation from the living state and yet permitted maximum resolution of the various [tissue] components.' The living state, unfortunately, hardly allows tissue being cut up into slices of 3 or 4 μm thick, and it being made visible through a microscope. Minimum deviation from the living state is achieved by 'interrupting the dynamic processes of the cell as promptly as possible and to stabilize the structure': by freezing the material or by fixation with formalin, dehydration with ethanol and stabilisation with paraffin (although these procedures may introduce *fixation artefacts*). A maximum resolution is reached by staining: 'using various dyes to increase the contrast of the tissue components'. *'Hematoxylin and eosin'*-staining is most often used, although the familiar pink and purple colour reveals 'little else than the character of the nucleus and the extent of the cytoplasm'. Apart from different staining techniques, a range of (immuno)chemical procedures (PAS-reaction, Immuno-Fluorescence) and physical methods (light microscopy, dark field –, phase contrast–, interference–, polarizing–, fluorescence–, UV –, and electron microscopy (c.f. Geneser 1986); spectrophotometry, X-ray projection or contact printing) is available.

In emergencies, material can be processed in a 'lightning action'. This is the case when the patient is being operated upon, and the continuation of the operation depends on the pathological findings. In such a case, material is brought in, dissected and deep-frozen. From this frozen section, slides can be sliced for

FIGURE 1: the diagnostic department as a semi-black box with informational flow



microscopical examination. The conclusions are communicated to the operation room through telephone or intercom. The surgeon may resume the operation within sometimes less than twelve minutes. Frozen sections are afterwards prepared into 'normal' slides for verification and final reporting.

In cytology, cells are examined that are located in a fluid environment; material is collected through a punction or smear (Hormoz 1994). Material is subjected to macroscopical observation only occasionally, and (clearly) needs not be dissected any further. Filtering or centrifuging the material may be needed to derive a richer sample of residue/sediment. The cells are stained for microscopic examination.

In chapter 3 of her doctoral thesis, Ellen Verheijen (1997) elaborately describes the outcomes of a study on the working procedures at three Dutch pathology laboratories.

3.3 Reporting diagnoses

Although figure 1 shows that different recipients (with different information needs) receive reports, only one report is composed at the pathology laboratory. Ever since a national organisation manages the information systems in The Netherlands, these reports are fairly uniform nationwide. Reports exist of various fields:

- ◊ coded data about the examination (laboratory identification number, type of examination, report number, year and date),
- ◊ data about the patient (name and birth date, gender, place of birth and current residence);
- ◊ free-text fields on observations: description of clinical status, clinical query, nature of the material, macroscopical observations, microscopical observations and conclusions. Optional fields are: additional examination (e.g. after an alternative staining or marker test) and revised conclusion (idem).
- ◊ a summary of the diagnosis in restricted language.

Only after completion of this report, copies are sent to the various recipients. The key fields vary between the destinations. The physician will concentrate on the conclusions paragraph and often only skimread the rest of the descriptions.

The full descriptions, especially the microscopical observations, are more important for the patient record: an older report may not be needed that often, but when it is needed, it should better have a complete, detailed and correct description of the observations. Even if the tissue slide (or an image of it) is still available at a later time, a detailed description gives the only representation of the interpreted information as available at the time of examination. This is important for medical and perhaps also medicolegal reasons.

The coded summary is added for computerized storage and retrieval of cases. Large medical databases prove useful in clinical and epidemiological studies, e.g. to produce incidence rates or to select individual cases that exhibit certain

characteristics. For these purposes, medical coding languages have been developed that can represent diagnoses in a structured, formal way. Many pathology laboratories, including those in The Netherlands, have adopted a version of SNOMED (Systematized Nomenclature of Medicine) (Gantner 1979). The coded version is an abstraction of the diagnosis. Coding requires decisions to be made on which of the (sometimes not exactly concording) formal terms gives the best representation. It also is a generalisation that demands a decision on the level of detail that is appropriate. A good classification of a case is correct, complete, but also concise.

The SNOMED coding system. 3.4

The SNOMED coding system is a descendant of SNOP (Standard Nomenclature of Pathology). In SNOMED, each term from a restricted (medical) lexicon has a pendant in a formal coding system (e.g.: the word 'forehead' is linked with code TY0110). SNOMED is a 'multi-axial' coding system – i.e., several axes were identified so that a term is coded in the following dimensions (after Gantner 1979). The first character of a code term identifies the axis to which the term belongs:

T topography: part, organ or region of the body. It often forms the 'anchor category';
P procedure: represents the operation through which the material was taken from the patient (e.g. *puncture*, *needle biopsy*);
M morphology: describes tissue structures or changes (e.g. *atrophia*);
E etiology: describes 'causal agents' – external factors that cause changes in morphology (e.g. *bacterial infection, laser burn*);
F function: represents physiological processes (e.g. *hyperthermia, pregnancy*); and
D disease: represents a diagnosis, often following topographical, morphological, etiological and/or functional observations.

In the latest release of SNOMED (version III, or SNOMED International), the number of axes was expanded to 18 (Rothwell et al. 1993).

Each of the axes has a hierarchical character which is visible in the code. After the first character – the axis identifier – five characters [0..9, X, Y] represent the relations between terms and the depth of the hierarchy on which this relation is defined. For instance, the word 'head' (code TY0100) is the 'parent' of the term 'forehead' (TY0110), 'sibling' of 'neck' (TY0600) and 'child' of the more general concept 'head and neck' (TY0000) (see fig. 2).

PALGA 3.5

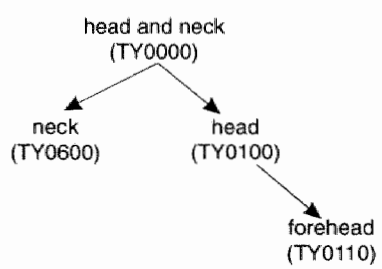
PALGA, as was stated before, means 'Pathologisch Anatomisch Landelijk Geautomatiseerd Archief' and stands for Dutch Network and National Database for Pathology. It is a department of SIG Health Care Information. The PALGA network was established by pathologists in 1971. Nationwide coverage of 100% (70 laboratories) was reached around 1990. The PALGA foundation manages a database in which an

excerpt of every examination is stored. Excerpts contain fields on the date and lab-ID of examination, (coded) patient data, requirements for follow-up examination and the encoded findings of the examination. In the latter field, terms in a restricted terminology are entered, which are then mapped to the formal SNOMED codes. Apart from maintaining the archive and network, PALGA gives technical support by installing and maintaining information systems in the laboratories.

The national archive contains about 20,000,000 excerpts (1996), with an annual increase of about 2,000,000. About 45% of this concerns histopathology, about 40% concerns cervix cytology, about 15% is about other cytology, and about 0.5% being obductions. The database is queried for patient data about 1,800,000 times per year; 30 to 40 times per year the database is used for data requests on diseases for scientific research. A number of publications that used 'PALGA-data' was cited in (PALGA 1992, 1995).

PALGA maintains close relationships with NVVP: Netherlandse Vereniging voor Pathologie (Netherlands Association of Pathologists).

FIGURE 2: example of hierarchical relations between SNOMED concepts



References:

Aller R.D.: Information Management. In: J.B. Henry (ed): Clinical Diagnostics and Management - 18th edition. Saunders Philadelphia PA 1991

Bauman R.A.: Reporting and communications. In: R.A. Greenes and R.A. Bauman (eds): Imaging and information management: computer systems for a changing health care environment; Special issue of The Radiologic Clinics of North America, Saunders Philadelphia PA, Vol. 34.3 1996.

Berman J.J. and Moore G.W.: SNOMED-encoded surgical pathology databases: a tool for epidemiologic investigation. Modern Pathology 1996 Vol 9 pp 944-950.

Bloom W. and Fawcett D.W.: A Textbook of Histology, 10th edition. Saunders Philadelphia PA 1975.

Chute C.G., Cohn S.P., Campbell K.E., Oliver D.E., and Campbell J.R.: The Content Coverage of Clinical Classifications. JAMIA 1996 Vol 3 pp 224-233.

De Bruijn L.M., Verheijen E., Hasman A., Van Nes F.L. and Arends J.W.: Speech interfacing for diagnosis reporting systems. Comp Meth Prog Biomed 1995 Vol 48 pp 151-156

Gantner G.E., Côté R. and Beckett R.S.: Systematized Nomenclature of Medicine Coding Manual, College of Americal Pathologists, 1979.

Geneser F.: Textbook of Histology. Munksgard Copenhagen 1986.

Griffith J.R.: The Well-Managed Community Hospital. Health Administration Press, Ann Arbor MI 1987

Hall P.A. and Lemoine N.R.: Comparison of manual data coding errors in two hospitals. J Clin Pathol 1986 Vol 39 pp 622-626

Hormoz E.: Cytology. In: E. Rubin and J.L. Farber: Pathology, 2nd edition. J.B. Lippincott Philadelphia PA 1994.

Martin B.G., Moore B. and McLendon M.W.: Organisation and management of the clinical laboratory. In: J.B. Henry (ed): Clinical Diagnostics and Management – 18th edition. Saunders Philadelphia PA 1991

PALGA: Jaarverslag (annual report, in Dutch), SIG Amsterdam, 1992.

PALGA: Jaarverslag (annual report, in Dutch), SIG Utrecht, 1995.

Rothwell D.J., Côté R., Brochu L.: The Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) International Microglossary for Pathology. College of American Pathologists, Northfield IL 1993.

Verheijen E.J.A.: Speech Technology for Medical Reporting: Consequences for the reporting process. Thesis Eindhoven University of Technology, 1997.

Woo J. and Henry J.B.: Clinical pathology / laboratory medicine purposes and practice. In: J.B. Henry (ed): Clinical Diagnostics and Management – 18th edition. Saunders Philadelphia PA 1991

CHAPTER · 2

MEDICAL LANGUAGE PROCESSING AND CLINICAL CODING: LITERATURE REVIEW

2 MEDICAL LANGUAGE PROCESSING AND CLINICAL CODING: LITERATURE REVIEW

This chapter is the result of a literature study that addressed the topics around processing of text in order to automatically assign classification codes to free text diagnosis reports. It appeared that medical informatics has been flirting with natural language technology for a long time, and for a broad range of objectives. Therefore, this review starts with a broad overview, after which our attention is zoomed in and focussed on encoding of pathology diagnoses. A number of projects is discussed that addressed natural language processing in close connection with classification. Other projects that concentrated on indexing medical literature are also discussed, because the methods from library science are closer related to the method that takes a central position in my research. The discussion summarizes the approach and results from the several projects, and explores the scientific and philosophical foundations of report classification.

1. Medicine and natural language

Chapter 1 showed that pathology diagnoses are primarily reported in natural language. Indeed, natural language plays a very important role in medicine in general, in scientific and clinical contexts, for communication and recording. Literature on the use of natural language in medicine is amply available. A query to MedLine, searching for 'natural language processing', indicated more than 75,000 records. The field can be divided into a large number of diverging topics - here given with some examples of publications:

- ◊ nomenclature definition (NIH 1997; Gantner et al. 1979; SNOMED 1997)
- ◊ definition of controlled languages (Kiuchi and Kaihara 1995, Musen et al. 1995)
- ◊ definition of formal languages and of coding and classification systems: SNOMED (1997), ICD (WHO 1997), Read codes (Chisolm 1990), Cimino 1996.
- ◊ thesaurus building and maintenance (Lovis et al. 1995, Ananiadou 1988, 1995)
- ◊ automatic medical coding (Sager 1981, 1994, 1995; Moore and Berman 1994ab)
- ◊ automatic translation (Kiuchi and Kaihara 1995)
- ◊ morphologic analysis of medical language (Lovis et al. 1995)
- ◊ syntactic analysis of medical language (Sager 1987)
- ◊ semantic analysis of medical language
- ◊ formal representation of medical knowledge/ concepts (Baud et al. 1992a), GALEN (Rector et al. 1994)
- ◊ automatic natural language generation (Wagner et al. 1995)
- ◊ literature index term extraction (SAPHIRE 1997, Blois 1984)
- ◊ natural language computer interaction
- ◊ medical narrative storage and retrieval
- ◊ automatic abstract generation

Furthermore, natural language has close relationships with other medical informatics domains, such as the electronic patient record (Tange 1997), structured data entry (Kirby et al. 1996), speech recognition in medical narrative dictation (Verheijen 1997, De Bruijn et al. 1995) and human-computer dialogue (Wulfman et al. 1993), medical knowledge acquisition (Musen et al. 1995), expert systems and decision support, storage and retrieval of medical literature (Wiesman and Hasman 1996), and the laying down of consensus and medical protocols.

Although this chapter concentrates mainly on literature on automatic diagnosis classification, it is impossible to see this separate from other topics. In the vast majority of projects that address automatic classification issues, the study is part of an integral natural language modelling project. So before I plunge in the sea of publications, I explore the issue of pathology reporting and classification.

2. The use of the pathology report

A clinical diagnosis, such as the outcome of a pathology examination, not only has instantaneous impact on the (continuation of the) patient's treatment but may also play a role in decisions on the patient's medical state that are to be made in the farther future, or in quality assessment studies. The case may even be useful for clinical or epidemiological studies. In all these situations, the value of the case lies in a clear description of the findings.

'Clearness' has a definition that changes with the user of the data. A natural language report is an excellent carrier of information between diagnostician and physician. Natural language is flexible, has a high representation power and yet is easy to generate and interpret for the human parties (Kiuchi and Kaihara 1995). Baud et al. (1992b) write: 'The richness of patient to patient information on a text basis, when dealing with discharge letters and consultant reports, is unquestionably more promising than any other encoding methodology.'

Natural language is less well suited if the recipient wants to apply computer logic to the outcome of the laboratory examination, or if the natural language that is used is simply not the language that the recipient has fluent command of. A practical solution for such communication bottlenecks was found in coding and classification systems: a formal coded description gives an abstraction of the case. In The Netherlands, each pathology report that is written is annotated by the pathologist with a classification in a SNOMED based system.

3. Classification and related concepts

Clear definitions are fruitful for this chapter, especially since the same words or terms are used for different concepts in different publications.

Starting with the physical world: any part of the perceivable world is an *object*. Abstraction on the basis of (sets of) objects forms a *concept* – a unit of thought or an idea. A *class* is an empirically or synthetically established group of objects.

The definition of a system of classes is named *taxonomy*. This refers both to the process of establishing the system as to the result of the process. The activity of assigning an object to a class is named *classification*, which is also used to refer to the result of that activity, or the resulting class identifier for that object.

A *code* is a formal representation of a class identifier. The activity of transcribing a class identifier to a formal representation is named *coding* or *encoding*. For this, a structured system of codes is necessary which may be named *coding system*, a product that is established after an activity that may be named *codification*. The meta-level of encoding plays no further role here.

Natural language concentrate around *words*. *Lexicography* collects words, and systemizes the collection into a *lexicon*. The contents of a lexicon is called the *vocabulary*.

Similarly, a restricted language for a domain uses *terms*, which are word-based but strictly defined identifiers of concepts. In a *nomenclature* process, terms are collected or defined; the product of this is also called *nomenclature*. The contents of a nomenclature is the *terminology*.

Classification is aiming for a situation where similar objects are grouped in specific clusters, so that in addressing a single cluster a whole group of objects can be processed in one instance. Classification can take place is two ways (cf. Watanabe 1985): (1) assigning a case to a certain predefined class (taxonomy), where each class has specific boundaries. This is a top down process that requires an initial definition phase in which classes and boundaries are set. (2) Forming (ad hoc) classes by grouping individuals that share a number of common features. It is a bottom up process that requires the recognition of the important features that contribute to a valuable grouping. This can only be defined in the context of the purpose – the reason why grouping is done in the first place (cf. Watanabe 1969).

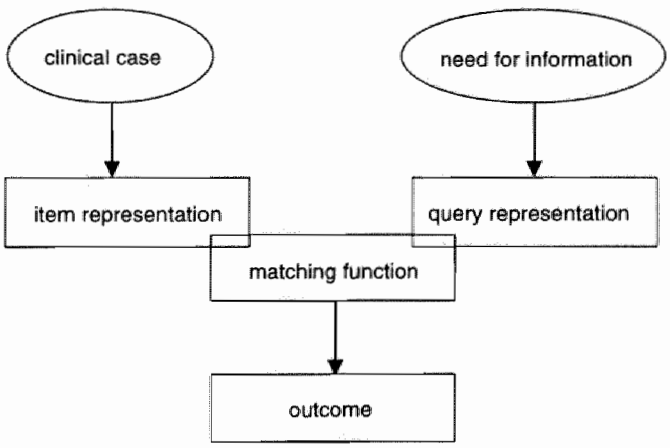
Cimino (1996) illustrates the importance of coding and classification to the medical informatics community: 'in the newly established Journal of the American Medical Informatics Association, 18 of the 51 papers in the first 8 issues deal with coding of clinical data. (..), in the most recent Symposium on Computer Applications in Medical Care (SCAMC 1995), of the 182 papers, 24 dealt specifically with controlled medical vocabularies, and an additional 65 dealt with applications requiring coded patient data'. Coding-languages are in constant development, both in breadth and in depth: they cover larger domains and are able to describe detail to a finer grain. This means that a coding system is not an inert adagium, but in using such a system one must anticipate updates, modifications and extensions.

4. Classification for retrieval

A formal classification seems to have its primary merit in retrieving cases from an archive, for whatever purpose. In this regard, the general Information Retrieval model can give some clarification (c.f. Ingwersen 1992): a matching function compares a certain query representation– which is an abstraction or formalisation of a user's need for information – to the stored representation of candidate instances – which are abstractions or formalisations of clinical cases as interpreted and translated by an expert (figure 1). It is clear that in a number of places in the graph, interpretations, decisions on categories or perhaps even errors contaminate the search and add noise to the output. Not all steps are even represented in the graph: for instance, a clinical case will give a conceptual representation in the observer's head first, partially depending on the technical circumstances such as imaging method, staining of slides, cutting artefacts etc. Then, this representation may lead to an abstraction which can lead to the phrasing in the natural language report. The classification of the case (the item representation in the database) is sometimes based on the report alone – if for example, this is independently done by a professional human encoder, or by a machine algorithm. Similarly, the distance between the user's need for information and the final query representation may cross several stations.

The quality of the output is usually given in the proportion-measures *precision* and *recall* (c.f. Van Rijsbergen 1979), where 'precision' gives an answer to the question 'are all items in the output relevant items?' – equivalent to the medical concepts *positive predictive value* – and *recall* gives an answer to 'are all relevant items retrieved from the collection?' – equal to *sensitivity*.

FIGURE 1: general model of information retrieval



In this circumstance, the classification is primarily an indexing function that directs the user/searcher to the full instance of the case, e.g. the free text narrative. Instead of attacking the problems of (manual) coding, one may also try to concentrate on retrieval with other mechanisms than the classification codes, e.g. full-text retrieval (Salton 1991, Hersh and Hickam 1995) or 'lightning hypertext' (Pathology Inc. 1997, Baak 1993).

In our situation, the endpoints were considered absolute: medical narrative was to be maintained in order to guarantee the full power of a fine-grained description language, and SNOMED coding was to be maintained in order to comply to international standards and allow formalisation of a case.

5. Coding and classification systems

The coding system that is used in The Netherlands is based on SNOMED, the Systematized Nomenclature of Medicine, successor of SNOP – Standard Nomenclature of Pathology (Gantner et al. 1979).

SNOMED in fact moves about in several separate fields (c.f. Read et al. 1995):

- ◊ terminology – each concept that is incorporated in SNOMED has one or more terms referring to it;
- ◊ encoding – each concept in SNOMED has a 6-character code attached
- ◊ grouping – cases can be retrieved with the use of terms or formal codes.
- ◊ structuring – the hierarchical character of the concepts, which is expressed in the structure of the SNOMED codes, include to a certain extent a semantical model of the medical domain.

SNOMED is a commercial system that is copyrighted, and charges fees for user licences.

Coding and classification of a case serves several goals: reduction of data, comparability of data (also in a plurilangual community), standardization and coupling with decision support systems (Van Bemmel et al. 1996)

The two main competitors for SNOMED are ICD and Read codes. ICD stands for International Classification of Diseases, and it has a long history that starts in 1893 with Jacques Bertillion's International List of Causes of Death (cf. Collen 1995). ICD is now in use in two versions: ICD-9-CM, the Clinical Modification of version-9 that was made by the Commission on Professional and Hospital Activities (ICD-9-CM 1979); and ICD-10, the tenth revision that is maintained by the World Health Organisation (WHO 1992).

The Read coding system is a British classification system (taxonomy) that was devised by James Read and adopted by the National Health Service (Chisolm 1990). The Read codes form a nomenclature where each term is annotated with a code.

GALEN is the name of a large European project in which a consortium is constructing a semantically sound model of clinical terminology – the GALEN Coding reference (CORE) model (Rector et al. 1994, GALEN 1997). Although the tasks of GALEN reach far beyond coding and classification, it aims to replace conventional coding systems with a concept based nomenclature. The GALEN coding reference system makes extensive use of relations between concepts to give a semantically sound representation of the case. GALEN is a project in full development, and is as yet not in use in established practical settings. The use of it in automatic text analysis is discussed later on.

SNOMED, ICD and Read codes are all used in practice for a longer period of time now. It frequently occurs that a single laboratory uses two classification systems in parallel, e.g. SNOMED for precise description of cases for pathology purposes, and ICD in order to accord with the hospital's information system, with national regulations or to feed insurance administation.

Chute et al. (1996) compared seven coding systems for expressiveness: ICD-9-CM, ICD-10, Read codes version 2.2, SNOMED version 3, UMLS version 1.3, CPT version 4 (Current Procedural Terminology – used mainly for payment purposes (cf. Collen 1995)) and NANDA (North American Nurses Diagnosis Association). Fifty clinical notes were collected, totalling 14.247 words that were parsed in 3061 medical concepts by a reviewer. These concepts were manually coded in each system. SNOMED stood out in the evaluation as 'a system with a broader coverage than any other.' On a scale of 0 to 2, SNOMED scored 1.74 overall, and 1.90 for the diagnosis component.

6. Problems with formal classification

The problems with the SNOMED based classification are twofold: (1) there are problems in the process of classification: it is a difficult task that involves deciding to which courser-grained class a fine-grained observation should belong, and that requires mastering the syntax and the allowed vocabulary. It is often necessary to look up classification terms in an extended code book; (2) the result is not optimal. The coding language still allows subjective variation between encoders, opinions differ especially on the level of abstraction and detail that the classification should have. Furthermore, the coding system has as yet no possibility to test for semantical inconsistencies in the classification.

In literature, similar bottlenecks are reported. Coles and Slavin (1976) report a 6.3% error rate for a human encoder on 99 single sentence summaries for pathology reports. Enlander reported 76% correct coding with manual data coding in SNOP terminology by residents at Stanford University Hospital, and Dinwoody and Howell report 12% to 17% error rates in coding hospital discharge data and general practice reports by professional encoders (see Hall and Lemoine 1986).

In a large scale study by Hall and Lemoine (1986), the production of the departments of morbid anatomy in two London hospitals was scrutinized for 2500 consecutive cases each. In one hospital, coding in SNOP, 16.1% of the coded reports were considered incorrect. In the other hospital, coding in SNOMED, 10.4% of the coded reports were erroneous. The majority of these errors (76%) was so large that cases would be irretrievable 'for a reasonably intelligent researcher in the future'. Secretaries were found to make errors in less than 1% of the cases, and errors due to poor handwriting were also marginal (1.1%). The authors state that 'many of the errors seem to be due to laziness in coding, with failure to consult the appropriate manual and reliance on memory for common codes.' Inconsistent coding was noted for identical reports by the same person. The authors write: 'many staff at both hospitals failed to appreciate the implications of poor or incomplete coding, and considered it to be a burdensome task. (...) We feel that the problems of manual coding are such that automatic computer encoding is the optimal method'. It was suggested that the higher error rate in one hospital was largely due to deficiencies in the SNOP system while easier coding was possible in the larger and more sophisticated SNOMED system.

Moore and Berman (1994b) write: 'The relative value of a coding language depends on the intended purpose of the coded database. Apparently, no single strategy for coding has been chosen by pathologists so that some pathologists code a single 'best fit' diagnosis for a lesion whereas others ensure lesion inclusion by coding multiple related terms. When administrators and epidemiologists attempt to use collected code data bases, they will have to contend with the diversity of coding approaches used by different pathologists.'

Carter et al. (1996) apparently support this view: 'Pathologist's failure to employ a consistent coding strategy is a major problem. Some pathologists employ a single 'best fit' approach, while others code with multiple related terms or modifiers.'

Chute et al (1996) wrote: 'Overall, these |clinical classification| systems fail to capture a large amount of clinical information (..). The capacity of these systems to support clinical decision making, facilitate outcomes research, or communicate the process of patient care seems limited.'

The late professor Wingert, editor of the German translation of SNOMED, advocated automatic support in classification and has presented significant research on the topic himself. He writes (Wingert 1986): '[the system's] main objective is not the correct indexing of each possible medical utterance but to free human indexers from indexing routine data, a job which is burdened with a high error rate, inconsistent indexing, and the practical impossibility of a retrospective verification of the indexed data. Therefore, the main policy adopted is to present a list of indexing proposals ordered according to a similarity function. Ideally, the list contains just the correct index. The human indexer may verify the proposal or select the correct index from the list.'

Research on medical narrative classification 7.

The problems with 'manual coding' have led to several research projects in which automatic coding of the case through the natural language report is tried.

Linguistic String Project – Medical Language Processing 7.1

The New York University group of Sager et al. has been researching natural language processing in medicine for more than 25 years now. The Linguistic String Project (LSP) (Sager et al. 1987) is a continuation of the work of Leo Tick and Julius Korein in the 1960s (cf. Collen 1995). When the research came to concentrate on medical language the name was appended with 'Medical Language Processing' (LSP-MLP). In their linguistic method, medical text is observed for regularities and word co-occurrences so that the semantical categories of words can be determined. The linguistic categories are ordered in columns such that a row in the (relational database) table can be read as a clinical statement, including:

> 'the finding *[column]* in the physiological function or the anatomic site of the patient *[col.]*, to the amount *[col.]*, found by procedure or laboratory test *[col.]*, was treated by *[col.]*; it occurred at time *[col.]* and/or with temporal features *[col.]*; it did not occur or there was some doubt *[col.]*.' (Sager et al. 1995).

Data in the table – especially that data that is stored in a single row – may be transferred to a coding language such as SNOMED. The five steps that are taken in processing a string are: (1) syntactic parsing, (2) resolving ambiguity and semantic labelling of the parse tree structures, (3) regularizing the parse tree, (4) converting the connective structure into a standard notation, and (5) linking the semantically labeled nodes of the final sentence tree with the corresponding nodes of the medical representation structure (Sager et al. 1993). Note that medical 'knowledge' is added in a relatively late stage of the process.

The LSP-MLP techniques were developed for English, but have been ported to French and implementation in German was reported to have begun (Nhan et al. 1989). These activities were employed in collaboration with the Geneva-project, which will be discussed later on.

Sager (1982) describes an experiment on reportings of early head/neck cancer symptoms: 150 free-text sentences were automatically classified to SNOMED terminology. Automatic coding comprised two stages: linguistic analysis with the LSP processor, and assignment of appropriate codes with a dedicated subprogram. 25 Cases did not reach the coding stage because the use of punctuation or parentheses differed from the modelled grammar (11) or parsing or formatting of the text failed (14). In 107 cases, the machine-retrieved codes matched those of manual coding, in 18 cases the codes did not match. It was concluded that 'a substantial portion of human coding time could be saved without lowering the quality of the data by the use of the automatic encoding procedure'.

In another experiment (Sager et al. 1993), hospital discharge summaries of asthma patients were abstracted in table structures that were developed by experts. The columns represented checklist values in six major categories of information, and thirteen categories in all. A total of 59 texts was processed with the LSP methods. Querying the table with SQL statements gave an average precision of 85.9% with 84.2% recall. If only major commissions and omissions were assessed, then precision would be 98.3% with 91.0% recall. The authors concluded: 'The semantic structuring and relative completeness of retrieved data suggest their potential use as input to further quality assurance procedures'. Note that in this study, the abstraction of classification systems such as SNOMED was not incorporated, but rather a newly specified structure was used.

Sager et al. (1995) used 20 case reports with patient discharge summaries and nurse notes. After automatic processing with the LSP methods, it was found that statements in the patient reports were typically clustered in three separate groups, namely statements on patient state, laboratory findings and treatment. Each of those may carry time or chronicity clauses, and doubt or negation arguments. SNOMED codes are sought by proceeding through the relational table row by row and performing string matching with SNOMED terms for the whole row, the separate fields, strings within the fields and separate words – whichever SNOMED term covers the most words in the least number of codes. Although the texts were also annotated by experts (in another experiment exercise), no exact assessment was published with regards to the system's performance.

7.2 RECIT / HELIOS

In 1987, the Geneva Canton University Hospital started a long term research program in Natural Language Processing (NLP) and Natural Language Understanding (NLU). The project is sometimes referred to as RECIT: Représentation du Contenu Informationnel des Textes médicaux. The Linguistic String Project was rewritten for the French language, and an alternative method – semantically-driven – was implemented. The NLP method was made part of the HELIOS project, which concentrated on the development of an integrated multimedia system. In HELIOS, natural language plays an important role in user-interaction, as a complement for other multimedia documents, or as a tool to unlock the vast quantity of narrative data that is stored (Engelsmann et al. 1994, Rassinoux et al. 1994). Furthermore, effort was undertaken to align the NLP/NLU method with the GALEN/GRAIL methods (Baud et al. 1993, Rassinoux et al. 1995).

The NLP module consists of three sub-components, which are (Rassinoux et al. 1994):

- ◊ the Analyser that processes free text sentences in English and French. The plurilingual character of the project demanded a language independent structure, which led to the adoption of a semantically driven approach. This contrasts with the largely syntactically driven LSP-MLP solutions. The choice for

semantical processing was argumented noting the more limited character of medical discourse over general natural language, and noting the relatively well-defined character of the medical domain (Rassinoux 1990, Baud et al. 1992a, 1992b). The analyser accepts statements in English and French and converts them into a single conceptual representation. Inclusion of German has been promised since 1992.

- ◊ the Dictionary Building Tools, which help to maintain the Medical and Linguistic Knowledge Base and extend it with concept types, dictionary entries, (hierarchical) relationships, proximity rules that also define relations, and so-called conceptual schemata.
- ◊ Queries on Conceptual Graphs - the Conceptual Graphs (CGs) being the knowledge representation. A query in natural language is parsed into a CG and matched with the stored CGs.

Baud and Rassinoux (1992b) propose a declarative style of programming the NLP system, where concepts are declared with the inclusion of appropriate features. Procedural processes, however, are also needed. The declarative style facilitates the construction of a pluri-lingual environment. The core of a system is a structured dictionary, that contains for each entry the syntactical characteristics, the semantical class to which it belongs, occurrence stamp, comments and pointers to typical examples. With a vocabulary of 20,000 words for a given medical specialty, and 200,000 to 300,000 words for the entire domain of medicine, this means that a huge task has to be considered, with adequate manpower resources (Baud et al. 1993). In 1993, a functioning system was described on the basis of a limited dictionary of 2000 words, in French and in English.

'As soon as the NLU tools have reached a good level of quality, (..) diagnosis encoding is possible using free text input, and automatic selection of near medical expressions in a given nomenclature. The final selection is performed under interactive control by the users.' (Baud et al. 1994). Michel, Lovis and Baud (1995) present a semi-automated encoding system for ICD-9, LUCID, that is however not structurally based but uses the lexical index term extraction technique. This electronic, context dependent code-book searcher relieves coding time and effort and improves precision and completeness (Michel et al. 1995). Unfortunately, no extensive results or technical details were published.

7.3 MENELAS

MENELAS (An Access System for Medical Records using Natural Language - Zweigenbaum 1994) is a natural language understanding system that uses Sowa's (1984) conceptual graph formalism. Delamarre et al. (1995) describe the use of Menelas for automated coding of patient discharge summaries. They consider this task as 'a compilation process of one specific language (the natural language) to a target language (the classification) with the help of an intermediate language'

(the conceptual formal representation). Texts were apparently processed per sentence - a morpho-syntactic analyzer produced sentence interpretations in accordance with language dependent syntax rules, a semantic analyzer built a conceptual graph representation, an finally a pragmatic analyzer processed the conceptual graph. A dedicated encoder synthesized the classification terms given the target system - in this case ICD-9-CM. The paper gives no specific results other than that the coding function of MENELAS showed good accordance to human coding.

At the university of Leuven, Dutch medical language is subjected to text analysis in a research partnership with the Menelas consortium (Spyns and De Moor 1996). The authors report gathering of 100,000 words in full form in their dictionary, which comes down to about 8,000 medical base forms. They decided on look-up instead of using morphological analysis; a category guesser handles the words that are unknown to the dictionary. The processing chain is complete: it runs morphologic lookup, and syntactic, semantic and pragmatic analysis, 'while integrating and reusing as much as possible available resources.'

The processor was tested on forty patient discharge summaries, containing 2253 Dutch sentences. Analysis was successful in 61.3%; about 10% of the sentences bounced on the parser's impossibility of handling conjunctions and disjunctions, and another 10% of the sentences was posed in a syntactically cripple manner.

7.4 GALEN

GALEN (Rector et al. 1994) is, apart from the name of the Greek pioneer in medicine, an acronym for Generalized Architecture for Languages, Encyclopedias and Nomenclatures in Medicine. It is set up as a semantically oriented nomenclature definition project. The European consortium that works on GALEN is co-ordinated from the University of Manchester, and makes part of the European Commission's Advanced Informatics in Medicine (AIM) programme.

GALEN aims to address the problems of clinical terminologies by constructing a semantically sound model of clinical terminology - the GALEN Coding reference (CORE) model. This model comprises

- ◊ elementary clinical concepts such as 'fracture', 'bone', 'left', and 'humerus';
- ◊ relationships such as 'fractures can occur in bones', that control how these concepts may be combined;
- ◊ complex concepts, e.g. 'fracture of the left humerus', made from simpler ones.

This compositional approach allows for detailed descriptions while preserving the structure provided by the individual components. A concept module, which should result in language independent concept representations, uses GRAIL: GALEN Representation and Integration Language. Coding Reference (CORE) models take care of the mapping of GALEN structures on existing coding structures.

The GALEN design is aimed at replacing conventional coding systems with a concept based terminology structure.

Aronow et al. 7.5

Aronow et al. (1995a, b) describe their efforts to disclose the text data in the automated medical record system of Harvard Community Health Plan. They describe an experiment in which specific asthma documents (those that describe asthma in acute exacerbation)were to be separated from other asthma documents. Two systems were available: (1) Inquery - a text based information retrieval system that presents the results in a rank order. Although classification is recognised by the authors to be a task differing from retrieval, Inquery's underlying model was estimated robust enough to provide good performance. Initial retrieval and expanded retrieval was assessed, where expanded retrieval meant that relevance feedback was introduced on the basis of a training set of positive and negative items. (2) Figleaf (Fine grained lexical analysis facility), a system under construction that bases classification on decision trees. Those decision trees are derived from examples in a set of training documents. The dictionary that was used, containing words and word-bigrams, was annotated by an expert. Results were presented in a ranked order.

Training data was derived from 75 asthma patient records (Aronow et al. 1995a), giving 988 encounter documents from which 293 encounters were assessed 'relevant' by an expert. A test data set was derived from 25 asthma patient records, giving 260 documents from which 60 were judged 'relevant'. The rank output allowed that precision and recall were presented in precision/recall operating curves. Initial Inquery performance was poor (.67 precision at .60 recall), expanded Inquery performed best (.87 precision at .80 recall) whereas Figleaf gave .77 precision at .80 recall.

The authors saw both techniques as 'quite promising', and would explore refinement in the Inquery's feedback mechanism, study on automating the training process for Figleaf and try to combine both systems in a 'serial-filter' setup (Aronow et al. 1995a). Figleaf would reduce the number of charts requiring manual review - which would allow cost savings and/or larger study sizes (Aronow et al. 1995b).

LBI 7.6

Brigl et al. (1994) propose Automatic Lexicon Based Indexing of Diagnoses in SNOMED (in short, LBI). Noun phrases in German with not more than 10 words are subjected to (1) pre-processing – the text string is divided into words, abbreviations are 'exploded' through table-lookup and word spelling is standardized; (2) morphological analysis – decomposition of words into word parts, determination of the lexemes and removal of irrelevant words; (3) semantic analysis – assignment of SNOMED codes to the set of lexemes with the use of the 'longest match' principle, and removing redundancy.

In (Brigl et al. 1995), the LBI method is put to the test, using 398 discharge diagnoses which were 'rather short formulated', and 385 medical report diagnoses

from pediatric surgery. All texts were in German. These texts originated from two pools of diagnoses: 8000 discharge diagnoses and 2400 medical report diagnoses. It is not clear whether these were the sets that aided in the construction of the lexicons.

In 54% of the discharge diagnoses, the automatically derived codes corresponded with the manually assigned codes; in 15% a courser set of codes was derived and in 10% the algorithms produced additional codes.

7.7 Several projects on term extraction

In a number of projects, search techniques were used for automatic classification and coding of texts. These techniques are typically used on short utterances such as single-line diagnosis summaries. Dictionary-lookup, or *indexing*, is used in order to find an exact match or the best match for the entire utterance. For each entry in the dictionary, a classification term or code is listed.

Coles and Slavin (1976) compared two computer programs, each used for coding 99 single sentence summaries for pathology reports. A word based SNOP encoder combined each topograpy term and each morphology term in the sentence and looked up a matching SNOP concept in the thesaurus. This gave an error rate of 17% for unprocessed sentences, but this error rate lowered to 6.3% after the sentences were processed manually such that several sentences with single pairs of topograpy and morphology were made. An alternative computer program initially failed to code 32.3% of the processed sentences, but this lowered to 3.3% after a second stage of summarisation by a consultant. This second summary was aimed at exact matching with the SNOP structure. As such, these systems offer little more than an automatic translation service from SNOP-terms to SNOP-codes.

Wingert (1985ab, 1986) described a methodology for automatic indexing in SNOMED. Part of his work was continued in the LBI-project (see section 7.6 of this chapter). Words are morphologically processed by omitting inflectional suffixes, segmenting compound words, and mapping orthographic variations (ph - f; ae - e). Connector words are omitted and the processed word-set is handled in an unstructured manner. Dictionary-lookup is done using the 'longest match' principle. Wingert (1986) reports no other results than: 'First tests have demonstrated that its rate of correct indexing is higher for human indexers if applied to a large set of routine data.'

Lin et al. (1992) describe a 'canonical phrase identification system' (CAPIS) that analyses physical examination findings in short phrases through concept based matching. Very little knowledge representation was incorporated in CAPIS. A phrase is entered into a lexical labeller that divides a character string into 'tokens', and tags those with syntactical labels. A simple heuristic parser analyses the text structure and forms a concept-word vector. This vector is then matched against a collection of predefined target vectors, checking for an exact match or the

best over-match (the target-vector is a subset of the test-vector) or finally the best under-match (v.v.). CAPIS was evaluated on a document collection from 20 patients with gastro-intestinal bleeding; for this collection an expert indicated a total of 156 findings. CAPIS automatically identified 144 of those, thus missing 12 findings, and falsely identified an additional 9 findings.

Carter et al. (1996) tested a commercial autocoder (PATHNET) for pathology phrases. The algorithm is proprietary, and not further discussed. It seems to be based on a personalised dictionary of phrase segments that is searched for an exact or best match when coding a new report. A 6-month monitoring period showed 631 errors in 2629 reports. Of these errors, 65% represented failure to code, even if the phrase was present in the dictionary; 5% were omissions in the dictionary that were considered important by the referee; 25% included an incorrect reference; and 5% of errors were codes that were included but meaningless. Adaption of the dictionary eliminated 10% of the errors (the second and the fourth of the error classes mentioned here).

SAPHIRE is an information retrieval front end to Medline; it allows input of queries in free text. SAPHIRE is claimed to provide 'potential solutions to the problems of inconsistent human indexing, clinician difficulties with controlled vocabularies and boolean searching, and the rich synonymy of medical language' (Hersh and Hickam 1995). SAPHIRE indexes the free text query into sets of full medical concepts, matches those concepts through the National Library of Medicine's *metathesaurus* to MeSH terms, searches the database, and ranks the output in order of relevance. In the experiment presented in (Hersh and Hickam 1993), SAPHIRE was compared with a boolean searching system (SWORD); 16 students submitted 10 queries to a database of 1992 documents, five with SAPHIRE and five with SWORD. SWORD gave better precision for queries 6-10 (71.1% vs. 52.6%). Both systems work equally well on recall (56.6% for SWORD and 57.6% for SAPHIRE), and precision for the queries 1–5 was 45.5% for SWORD and 44.3% for SAPHIRE. SAPHIRE can be seen in action as a search engine that finds MeSH terms from a natural language query (SAPHIRE 1997); it would be usable to extract index terms from summary lines.

8. Discussion

The above compilation shows that the topic of automatic classification of medical narrative is attacked in quite a number of projects. Many of these projects address automatic classification from a grand concept that concentrates around the representation of conceptual knowledge. A natural language utterance is analyzed, a conceptual representation is formed of what is described, possibly the concept representation is verified using knowledge structures, and finally a classification is generated. The methods are usually evaluated with experiments on short sentences.

The steps involved in natural language understanding in these projects are generally: (1) word level analysis: morphological analysis of the word and deriving its semantical characteristics (concept class, e.g. 'this word is a morphology identifier') or syntactical characteristics(e.g. 'verb, third person plural'); (2) syntactic analysis - mapping the words onto a valid structure or disambiguating the word string; (3) semantic analysis – constructing a meaning through, typically, a conceptual graph, and (4) handling the meaning structure depending on the objective of the study.

Research that uses such a structured analysis method shows two trends: one is to concentrate on mapping a language utterance onto a syntactical representation and to make inferences from that point on, the other is to aim for a sound semantical representation. The latter approach requires extensive modelling of the domain, something that GALEN is working on; because of its domain dependence, it can be made less dependent on the language that the text is written in. In syntactically oriented projects it was decided to adopt language theories that were available and commit themselves less to an elaborate semantical representation of the domain.

These systems may be compared to a solar system – it all revolves around a core model (illuminated or not), and more than one satellite can use its resources, but it takes a lot of energy to make it work. For instance, the structured dictionary of words and their characteristics and relations in the RECIT/HELIOS project was estimated to build up to 200,000 to 300,000 words, while the current state reports inclusion of 2,000 to 3,000 words. A significant part of the work in this mammoth project relies on manpower resources.

Syntax orientation – which is the predominant character of for instance the LSP-MLP system – requires language modelling rather than domain modelling and may profit from language research in other fields than medicine. The LSP system was originally developed for analysis of general English and was only later applied for medical text. The syntax oriented method is in theory less domain dependent than semantically oriented systems. LSP-MLP was indeed tested for texts from a number of different medical domains, but unfortunately within each experiment only a small, specified domain was addressed. No broad experiment was described.

Syntax oriented systems are language dependent, and support cross-language functions (e.g. automatic translation) to only a basic level. It is not surprising that the swing towards semantically oriented development, which is language independent and domain dependent, took mainly place in the polyglottal European community. Baud et al. (1992a) claim it would not be necessary to redefine the conceptual grammar of their RECIT/HELIOS system: 'The NLP system [...] would appear easily adaptable to any languages other than French.', however: 'Of course it would have to be rewritten in the context of each separate language. It may even have to be partly redesigned'.

The projects that were presented above use declarative statements to infer the structures, be they semantical or syntactical. This may be strange, since the originator branch of science of this approach – artificial intelligence – has virtually abandoned the idea. In the words of Eugene Charniak (1993): 'I think it's fair to say that few, if any, consider the traditional study of language from an artificial intelligence point of view a 'hot' area of research. (...); it is increasingly hard to believe that it will shed light on broader problems, since it has steadfastly refused to do so in the past.' Newer techniques which were advocated instead, such as construction of probabilistic context-free grammars (which use tree probabilities) and part-of-speech tagging with Markov chains (cf. Charniak 1993), were not as such found in the consulted literature.

A number of other projects was described that used index term extraction, a method that requires very little domain modelling but rather an extensive library of utterances and connected classification terms. Index term extraction can only give a good classification if short utterances (preferrably summarizing phrases) are used, because the number of classification terms that would be derived from a larger text would be superfluous.

Techniques in index term extraction originate from research in library science and information retrieval. Methods are mainly statistically oriented and rely on finding best matches with terms in a dictionary. A similar approach could be used for entire texts: a newly encountered text is classified with the help of an annotated collection of other texts. From the collection, the archive text that is the most similar to the new text is retrieved. This archive text gives a suitable classification if a number of conditions are met, which include (1) the collection contains a suitable classification (this is the 'coverage' of the collection, cf. Van Rijsbergen 1978); (2) the similarity between new text and archive text is sufficiently high; (3) the similarity measure is defined such that high values guarantee that the reports are also semantically very similar – and thus provide a suitable classification (which may be seen as the 'predictive value' of the similarity measure).

Note that

- ◊ most medical departments have assimilated a large archive of texts, manually annotated with clinical codes. This makes construction of a collection a fairly easy exercise;
- ◊ this method has no dependencies with respect to the language or coding system. These characteristics lie outside the algorithms, only in the collection;
- ◊ the method does not use inherent characteristics of the coding system, e.g. using the corresponding terminology for direct mapping;
- ◊ changes in the coding system and gradual shifts in word use require no adaptation of the method's algorithms. The collection is kept up-to-date with inclusions of recent reports and draining away the older ones; and
- ◊ the method does not rely on the availability of a domain model.

With this concept, the 'classification task' has become a 'search task'. The search of text items in large text collections is an activity that is addressed in the discipline 'Information Retrieval' (Salton & McGill 1983, Van Rijsbergen 1978), which can be regarded as a child of Library Science and Computer Science. Library science, of course, also has a long tradition of item cataloguing, classification and also coding. A revolution in the organisation of libraries was provoked by Melville Louis Kossuth Dewey with his publication 'A classification and subject index for cataloguing and arranging the books and pamphlets of a library' (Dewey 1876). His system is still known as Dewey Decimal Classification (DDC) and is the ancestor of a number of modern library systems, including a descendant version of DDC itself (Chan et al. 1996).

Evidently, automated systems for cataloguing or indexing large collections of diverse literature could never wait for a conceptual core model of the world to be constructed, and alternative methods were developed. These methods include automatic key word extraction by calculating word frequencies, up to automatic summarization and abstract generation.

Classification of a text – medical or other – means that we must understand the text. *Understanding* is more or less a synonym of *intelligence*, or more precisely: intelligence is the potential for showing understanding. So classification is a job for intelligence, be it artificial or natural. Intelligence is not something that is present or absent, it rather is a degree between 'stupid' and 'smart'. In his book The Cognitive Computer, Roger Schank presents an understanding-spectrum:

making sense	———	cognitive understanding	———	complete empathy

Schank (1984) writes: 'The endpoints of this spectrum can be described loosely as, on the one hand, the understander saying to herself, Yes, I see what is going on here, *I know what it is about* and, on the other hand, her saying My God, that's exactly what I would have done, I know precisely how you feel'.

A program for automatic classification needs only to say: 'I know what it is about', and does not require much more than a very low level of intelligence. It must not necessarily be able to explain how it came to that conclusion – an introspective quality. Whether or not such behaviour is an intelligent activity depends on the eye of the beholder, which is known to change over time (cf. Schank 1984, Weizenbaum 1976).

Schank's work on (artificial) intelligence concentrates on scripts. 'Scripts are prepackaged sets of expectations, inferences, and knowledge that are applied ... like a blueprint for action. ... A script tells what is likely to come next in a chain of events that are stereotypes'. Scripts also are vague enough to invoke response to events that are somewhat deviant from the original scenario. This in a way allows to act on new situations: generalize, predict, learn, and make inferences.

Winograd and Flores (1986) saw a shift in AI-research from 'the traditional problem-solving orientation towards a new one centered around 'frames' or 'expectations''. Problem solving shifts to recognition, understanding is treated as pattern recognition. Minsky (1975) writes: 'When one encounters a new situation (...) one selects from memory a substantial structure called a frame. This is a remembered framework to be adapted to fit reality by changing details as necessary'. Frames contain details that have default values to be altered if the perceived situation gives reason to do so.

In a later book – Tell Me A Story – Schank (1990) came to use an even less formal concept of scripts: stories. Rather than rule based, Schank sees human memory to be 'story-based', and intelligence is the 'apt use of experience' (stories). Understanding then means searching memory. His initial definition that 'intelligence is really about understanding what has happened well enough to be able to predict when it might happen again' consequently translates into 'intelligence is really about searching memory for suitable stories, so that the consequenses of the present situation can be predicted from the course of the past stories.'

Story-based intelligence depends on three factors. (1): The collection of memory items determines if scripts are available to be applied in a situation. A tourist in a new country will easily make a social blunder because of the lack of suitable scripts. Learning means growing and pollarding memory: storing new scripts, newer versions of scripts, or different branches of existing scripts, and generalising scripts by combining them. (2): The selection of the appropriate script. The first factor showed that memory is ideally very large and dynamic. It requires a grand capacity of navigating this memory – selecting and rejecting candidate-scripts. This factor also includes the handling of vague information: seeing the similarity and discrepancy between differing scenes and scenarios. (3): The probabilities that certain results follow from a certain sequence of events. A scene may take a different turn than you expected. As such, the perceived outcome feeds the learning process.

These factors – having a memory, searching that memory and using the findings to generate an output, are the same key elements as in the classification method that was sketched at the beginning of this chapter.

The shift from rule-based intelligence to recognition based intelligence means a shift from cognition to recognition, from logical to analogical, from deduction to induction, from dichotomy to probability...
from knowing to guessing.

But.. The adoption of one basis for intelligence is not a realistic one. 'What is stored in memory is not necessarily synonymous with what a program knows'. A computer may have 'the memory of an elephant', but merely storing numbers is far from

representing knowledge (Hofstadter 1979). It is essential to be able to retrieve items from memory, but even then, 'real thinking is an interplay of deduction and induction' (Watanabe 1985). Artifical intelligence research may expect better results if script based methods are combined with rule-based knowledge statements (cf. Croft 1993).

9. Conclusions

Research on automatic classification of medical narrative is ususally found within larger projects that address a wide span of medical informatics issues. Traditional coding and classification may be abolished when there is the technical possibility to make new conceptual representations of medical data that support document storage and retrieval – and that may also allow, for instance, easy translation of clinical data between languages. However, this is a utopian view for the time being.

The current paradigm in automatic classification involves pursuing a high level of natural language understanding or domain knowledge. Emphasis is laid on either syntactical or semantical modelling, although adequate results are apparently not possible without crossing this border at one point in the process. The approaches have a character of logic and deduction, although probabilistic elements are now and then encountered – notably category guessing of newly encountered words and 'proximity processing' of conceptual graphs.

The disadvantage of attempting to reach a high level of natural language understanding or domain knowledge is that one risks to end up with a rigid system that needs constant manual maintenance in order to keep up with domain developments and with the human's natural inventiveness for thinking up original utterances. Moreover, projects such as these need a large scale setup.

With these observations considered, an alternative paradigm was explored: one that is rather based on recognition, induction and probability. Evidence for this approach can be found in literature on Information Retrieval, Pattern Recognition, and Natural Language Processing. Studies in a medical environment with this approach were found, but they concentrate on short utterances (typically summaries). Studies with larger, unstructured texts (for instance complete reports) have not been reported.

The central question of this thesis, that was phrased before as: *'Is it possible to automatically assign classification codes to a diagnosis report that is written in natural language'* is refined on the basis of the observations from literature. Extensive domain modelling of pathology is confronted with practical limits of the project, and no ready-to-use models are available. Classification methods using language models were described in literature, but were found to be applied on limited domains only. Again, extensive modelling would be necessary to cover the wide

range of pathology utterances that can be found in the reporting routine. Model-based systems are prone to be rigid, so that effort would start almost from scratch when new domains were to be entered. Extension of the domain would require adaptation of the processing routines, probably by hand. Model-based methods were therefore discarded from consideration. Instead, attention was given to recognition based methods because these would be more challenging from a scientific point of view. Nearest neighbor classification of texts through vectors in the vocabulary space was explored and applied to pathology reports. The operationalised research question then has become: *'Can recognition-based techniques classify diagnosis reports that are written in natural language; how well would such techniques perform, and where lie their weaker points?'*

References

Anadiadou S.: Towards a methodology for Automatic Term Recognition. PhD Thesis, UMIST Manchester, 1988.

Anadiadou S.: Automatic recognition of medical terminology. In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 3-7.

Aronow D.B., Soderland S., Ponte J.M. et al.: Automated classification of encounter notes in a computer based medical record. In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 8-12 (1995a)

Aronow D.B., Cooley J.R., and Soderland S.: Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record. Proc Annu Symp Comput Appl Med Care 1989 pp. 309-313 (1995b)

Baak J.F.A.: The Lightning hypertext of cytology, software review, Ned Tijdschr Geneeskd 1993 Vol 137 p 474

Baud R.H., Rassinoux A.M., and Scherrer J.R.: Natural language processing and semantical representation of medical texts., Methods Inf Med 1992 Vol 31 pp 117-125 (1992a)

Baud R.H., Rassinoux A.M. and Scherrer J.R.: Natural Language Processing and Medical Records. In: Lun K.C. et. al (eds). MEDINFO 92. Amsterdam North Holland 1992 pp. 1362-1367. (1992b)

Baud R.H., Lovis C., Alpay L., Rassinoux A.M., Scherrer J.R., Nowlan A., and Rector A.: Modelling for Natural Language Understanding. In: Safran C. (ed). Proc Annu Symp Comput Appl Med Care 1993 pp. 289-93.

Baud R.H., Alpay L., and Lovis C.: Let's meet the users with Natural Language Understanding. In: Proc. EPISTO AIM working Conference on KB systems, Munich 1993. Amsterdam: IOS Press, 1994.

Berman J.J., Moore G.W., Donnelly W.H., Massey, J.K., and Craig B.: SNOMED Analysis of 40,124 Surgical Pathology Cases. Am J Clin Pathol 1994 Vol. 102 pp 539-540

Blois M.S.: Information and medicine: the nature of medical descriptions. University of California Press, Berkeley 1984.

Brigl B., Mieth M., Haux R., and Glück E.: The LBI method for automated indexing of dianoses by using SNOMED: Part 1: design and implementation. Int. J. Biomed. Comput. 1994 Vol 37 pp 237-247

Brigl B., Mieth M., Haux R., and Glück E.: The LBI method for automated indexing of dianoses by using SNOMED: Part 2: Evaluation. Int. J. Biomed. Comput. 1995 Vol 38 pp 101-108

Carter K.J., Rinehart S., Kessler E. et al.: Quality assurance in anatomic pathology: automated SNOMED coding. JAMIA 1996 Vol 3 pp 270-272.

Charniak E.: Statistical Language Learning. Bradford/MIT press 1993.

Chute C.G., Cohn S.P., Campbell K.E., Oliver D.E., and Campbell J.R.: The Content Coverage of Clinical Classifications. JAMIA 1996 Vol 3 pp 224-233.

Cimino J.J.: Review paper: coding systems in health care. Methods Inf Med 1996 Vol 35 pp 273-284

Chisolm J.: The Read Classification. British Medical Journal 1990 Vol 300 (6732) p 1092.

Coles E.C. and Slavin G.: An evaluation of automatic coding of surgical pathology reports. J.Clin. Path., 1976, Vol 29 pp 621-625

Collen F.M.: A history of Medical Informatics in the United States - 1950 to 1990. American Medical Informatics Association 1995.

Croft W.B.: Knowledge-based and statistical approaches to text retrieval. IEEE Expert 1993 p 8-12

De Bruijn L.M., Verheijen E., Hasman A., Van Nes F.L., and Arends J.W: Speech Interfacing for Diagnosis Reporting Systems: an overview. Comp Meth Prog Biomed 1995 Vol 8 pp 151-156.

Delamarre D., Burgun A., Seka L.P., and Le Beux P.: Automated coding of patient discharge summaries using conceptual graphs., Methods Inf Med 1995 Vol 34 pp 345-351

Dewey M.L.K.: A classification and subject index for cataloguing and arranging the books and pamphlets of a library. Amherst MA 1876.

Engelsmann U., Jean F.C., and Degoulet P. (eds): The HELIOS software engineering environment: Results of the European AIM project HELIOS-2. Supplement to Computer Methods and Programs in Biomedicine 1994 pp S1-S152.

GALEN: http://www.cs.man.ac.uk/mig/galen/brochure.html (1997)

Gantner G.E., Côté R., and Beckett R.S.: Systematized Nomenclature of Medicine Coding Manual, College of Americal Pathologists, Northfield 1979.

Hall P.A. and Lemoine N.R.: Comparison of manual data coding errors in two hospitals. J. clin. Pathol., 1986 Vol 39 pp 622-626

Hersh W.R. and Hickam D.H.: A comparison of two methods for indexing and retrieval from a full-text medical database. Medical Decision Making 1993 Vol 13 pp 220-226.

Hersh W. and Hickam D.: Information retrieval in medicine: the SAPHIRE experience. In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 1433-1437.

Hofstadter D.R.: Gödel, Escher, Bach: An Eternal Golden Braid. Basic Books, New York 1979.

ICD-9-CM: Commission on Professional and Hospital Activities: International Classification of Diseases ninth revision with Clinical Modifications. Ann Arbor 1979.

Ingwersen P.: Information Retrieval Interaction. Taylor Graham, London 1992

Kirby J., Cope N., De Souza A. et al.: The PEN&PAD data entry system. In: J. Brender et al. (eds): Medical Informatics Europe 1996, IOS press Amsterdam pp 430-434

Kiuchi T. and Kaihara S.: Machine translation and medical linguistics. Medinfo 1992 pp 1350 -

Kiuchi T. and Kaihara S.: On the linguistic representation of medical information: natural language, controlled language, and formal language. In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 23-27.

Lin R., Lenert L., Middleton B., and Shiffman S.: A free-text processing system to capture physical findings: Canonical Phrase Indentification System (CAPIS). Proc Annu Symp Comput Appl Med Care 1992 pp. 168-172

Lovis C., Michel P.A., Baud R., and Scherrer J-R.: Word segmentation processing: a way to exponentially extend medical dictionaries. In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 28-32.

Michel P.A.., Lovis C. and Baud R.: LUCID: a semi-automated ICD-9 encoding system. In: R.A. Greenes et al.: MEDINFO 1995 p 1656.

Miller E.T., Wieckert K.E., Fagan L.M., and Musen M.A.: The development of a controlled medical terminology: identification, collaboration, and customization. In: R.A. Greenes et al. (eds): Medinfo 1995 pp 148-152

Moore G.W. and Berman J.J.: Automatic SNOMED coding., Proc Annu Symp Comput Appl Med Care 1994 pp 225-229 (1994a)

Moore G.W. and Berman J.J.: Performance analysis of manual and automated Systemized Nomenclature of Medicine coding. American Journal of Clinical Pathology 1994 Vol 101 pp 253-256 (1994b)

Musen M.A., Wieckert K.E., Miller E.T. et al.: Development of a controlled medical terminology: knowledge acquisition and knowledge representation. Meth Inform Med 1995 Vol 34 pp 85-95

Nhan N.T., Sager N., Lyman M., Tick L.J., Borst F., and Su Y.: A Medical Language Processor for Two Indo-European Languages. Proc Annu Symp Comput Appl Med Care 1989 pp. 554-558.

NIH: http://www.nlm.nih.gov/research/umls/umlsdoc.html (1997)

Pathology Informatics Inc.: http://www.pathinfo.com/ (1997)

Rassinoux A.M., Michel P.A., Juge C., Baud R., and Scherrer J.R.: Natural language processing of medical texts within the HELIOS environment. Comput Methods Programs Biomed 45 Suppl: S79-96, Dec, 1994.

Rassinoux A.M., Wagner J.C., Lovis C., Baud R.H., Rector A., and Scherrer J.R.: Analysis of medical texts based on a sound medical model. Proc Annu Symp Comput Appl Med Care 1995 pp 27-31

Read J.D., Sanderson H.F., and Drennan Y.M.: Terming, encoding and grouping. In: R.A. Greenes et al. (eds): Medinfo 1995 pp 56-59

Rector A.L., Nowlan W.A., and the GALEN consortium: The GALEN project. Cop Meth Prog Biomed 1994 Vol 45 pp 75-78

Sager N., Bross I.D., Story G., Bastedo P., Marsh E., and Shedd D.: Automatic encoding of clinical narrative., Comput Biol Med 1982 Vol 12 pp 43-56

Sager N., Friedman C., and Lyman M.S.: Medical Language Processing: Computer Management of Narrative Data. Addison-Wesley, Reading MA. 1987

Sager N., Lyman M., Tick L.J., Nhan N.T., and Bucknall C.E.: Natural language processing of asthma discharge summaries for the monitoring of patient care. Proc Annu Symp Comput Appl Med Care 1993 pp 265-268.

Sager N., Lyman M., Nhan N.T., and Tick L.J.: Automatic encoding into SNOMED III: a preliminary investigation., Proc Annu Symp Comput Appl Med Care 1994 pp 230-234

Sager N., Lyman M., Nhan N.T., and Tick L.J.: Medical language processing: applications to patient data representation and automatic encoding. Meth. Inform. Med. 1995 Vol 34 pp 140-146.

Salton G. and McGill M.J.: Introduction to modern information retrieval. McGraw-Hill, New York 1983.

Salton G.: Developments in automatic text retrieval. Science 1991 Vol. 253 pp 974-980

SAPHIRE: http://www.ohsu.edu/cliniweb/ post-saphire (1997)

Schank R.C. with Childers P.: The Cognitive Computer: On Language, Learning, and Artificial Intelligence. Addison Wesley, Reading MS 1984.

Schank R.C.: Tell Me a Story: A New Look at Real and Artificial Memory. MacMillan, New York 1990.SNOMED: http://snomed.org/ (1997)

Sowa J.F.: Conceptual structures: Information processing in mind and machine. Addison Wesley New York 1984

Spyns P. and Willems J.L.: Dutch medical language processing: discussion of a prototype. In: R.A. Greenes et al. (eds): Medinfo 1995 pp 37-40

Spyns P.: Natural Language Processing in Medicine: An Overview. Meth Inform Med 1996 Vol 35 pp 285-301

Spyns P. and De Moor G.: A Dutch medical language processor. International journal of bio-medical computing 1996 Vol 41 pp 181-205

Tange H.: Medical Narratives in the Electronic Medical Record: towards a searching structure with optimal granularity. PhD thesis, Maastricht University 1997.

Van Bemmel J. (ed): Handbook of medical informatics - preliminary version. Bohn Stafleu van Loghum, Houten 1996

Van Rijsbergen C.J.: Information Retrieval (second edition). Butterworths, London 1979.

Verheijen E.J.A.: Speech Technology for Medical Reporting: Consequences for the reporting process. Thesis Eindhoven University of Technology, 1997.

Wagner J.C., Solomon W.D., Michel P.A., Juge C., Baud R.H., Rector A.L., and Scherrer J.R.: Multilingual natural language generation as part of a medical terminology server.In: R.A. Greenes et al. (eds): MEDINFO 1995 pp 100-104.

Watanabe S.: Knowing and Guessing: A Formal and Quantitative Study. Wiley New York 1969.

Watanabe S.: Pattern Recognition: Human and Mechanical. Wiley New York 1985.

Weizenbaum J.: Computer Power and Human Reason: from judgement to calculation. Freeman 1976, Pelican edition, Penguin 1984.

WHO - World Health Organisation: ICD-10 - International Statistical Classification of Diseases and Related Health Problems: 10th revision. Geneva 1992.

Wiesman F. and Hasman A.: A graphical user interface for biomedical literature search. In: J. Brender et al. (eds): Medical Informatics Europe 1996, IOS press Amsterdam pp 624-628

Wingert F.: Automated indexing based on SNOMED. Meth Inform Med 1985 Vol 24 pp 27-34 (1985a)

Wingert F.: Morphologic analysis of compound words. Meth Inform Med 1985 Vol 24 pp 155-162 (1985b)

Wingert F.: An indexing system for SNOMED. Meth Inform Med 1986 Vol 25 pp 22-30

Winograd T. and Flores F.: Understanding Computers and Cognition: A New Foundation for Design. Ablex, Norwood NJ 1986.

Wulfman C.E., Rue M., Lane C.D., Shortliffe E.H., and Fagan L.M.: Graphical access to medical expert systems: V. Integration with continuous-speech recognition. Meth. Inform. Medicine 1993 Vol 32 pp 33-46

CHAPTER · 3

NEAREST NEIGHBOR CLASSIFICATION: METHOD AND PILOT EXPERIMENT

Acctepted for publication in: International Journal of Technology Management as: De Bruijn L.M., Hasman A., Arends J.W.: Classification of diagnoses that are described in natural language.

3 NEAREST NEIGHBOR CLASSIFICATION: METHOD AND PILOT EXPERIMENT

1. Introduction

We present a method to automatically code or classify natural language narrative. In our research, we used diagnosis reports from a pathology laboratory. In such a laboratory, organic tissue or fluid is examined to verify the presence and absence of diseases. The diagnosis is reported to the requesting clinician (general practitioner, surgeon) through a report in natural language. A further report is made for the archiving organisation – in our country a national archive – that stores diagnosis excerpts for patient monitoring and for epidaemiological studies. For accurate storage and retrieval, a formal code is added to the report. This code summarises the description of the material and the diagnosis, and is composed in a medical coding system named SNOMED (Systematized Nomenclature of Medicine).

A SNOMED code gives a complete and systematic description of the diagnosis, including the source of the material, the disease that was found and, if necessary, sidedness and radicality. SNOMED coding sets the pathologist a difficult task, because the system is so rigid that a thesaurus must be consulted to find accurate and correct coding terms. On the other hand, the coding system is still loose enough to allow subjective preferences, and correctness and precision are not guaranteed. So both the archiving organisation and the pathology laboratories expressed the need for automatic support in SNOMED coding.

A practical use solution would, for example, be: after a natural language report is composed by the pathologist, the text is analysed and a formal classification is automatically derived. This is returned to the pathologist, who verifies it and may choose to request alternative classification terms or alter them manually. For satisfactory use, performance must be flawless for routine cases, but need not be perfect for difficult cases – if only because perfection is impossible to define in this task.

Note that the actual classification process is kept to the pathologist. It is more a coding problem that is solved: recoding a natural language diagnosis into a formal language.

The nearest neighbor method that is presented is not dependent on the language or domain of the natural language texts, nor on the coding system that is used. This makes it possible to apply the method to other languages or other domains, medical or non medical, to classify natural language text.

2. General method

The problem of classifying natural language texts can be attacked in two ways:
◊ with Natural Language Processing techniques, and
◊ with Pattern Recognition or Information Retrieval techniques.
In the Natural Language Processing (NLP) approach, a text is parsed into syntactical structures, and semantically analysed. Although controllability and precision are great advantages of the NLP method, it requires a large extent of domain and language dependent modelling (cf. Croft 1993).

An alternative approach was found in Pattern Recognition (PR) and Information Retrieval (IR) techniques. Although PR and IR are not historically linked, they arrive at almost the same solutions for text classification. The method in this paper is described along the lines of Pattern Recognition because of its clear and stable theoretical basis.

For classification, one represents the observed pattern (in our case the natural language report) in a defined vector space, and calculates to which predefined class the vector most probably belongs (Fukunaga 1993, Schalkoff 1992). An optimum result can be obtained when the distribution function per class over the vector space is computed. If it is not viable to calculate all the distribution functions – which is the case in our polyclass problem – one can choose to place a new sample under the class of the nearest of a set of previously classified points. This is called the Nearest Neighbor rule (Cover and Hart 1967). The rule can be extended to a number of near neighbors – a polling scheme then decides on the final result of this *k*-Nearest Neighbor rule.

In order to apply the Nearest Neighbor rule, one must have
◊ a set of previously classified points – in our case an archive of previously coded pathology reports;
◊ a distance metric that defines the nearness of two points – in our case the textual (dis)similarity between two pathology reports;
◊ the definition of a vector space in which this distance metric can be computed.
The space is a multidimensional vector space, in which one dimension is assigned to every feature that is recognised in the samples. The scale per dimension can represent a weight to stress the importance of that feature for classification.

The distance metric that is most commonly used, is in fact a similarity measure: the normalised inner product between two vectors (cf. Van Rijsbergen 1979):

$$d(X_1, X_2) = \sum_{i=1}^{D} x_{1i}x_{2i} \; / \; \left(\sum_{i=1}^{D} x_{1i}^2 \sum_{i=1}^{D} x_{2i}^2 \right)^{1/2} \qquad (1)$$

with x_{1i} is the value for dimension i in vector X_1 ; i is taken over space D.

This metric equals 1 for identical vectors, 0 for completely different vectors and a value in between for more or less similar vectors.

3. Defining the vector space

Using nearest neighbor pattern recognition for text matching requires specific decisions on feature selection, feature transformation and scaling.

3.1 Feature selection

Textual narrative data is composed of characters, forming words, ordered in a meaningful structure. This gives several different starting points for feature extraction. An obvious choice is using words for feature units. Words that are shared by two texts contribute to a higher similarity for that pair of texts. This will result in a high dimensionality for the vector space – an order of magnitude of about 20.000 to 50.000 in our application. The paragraph on feature transformation will show that the dimensionality can be significantly reduced.

Other text features that could be considered are sub word units (syllables), super word units (word sequences such as word bigrams or trigrams) or even units that are not based on word boundaries. An example of the last category is given by Damashek (1995), who uses n-grams (e.g. 5-grams: consecutive sequences of 5 characters). An advantage of Damashek's method is that texts in different languages, or written with different spelling preferences, may still share a number of n-grams. Its disadvantage is its rapidly growing dimensionality.

The use of words as feature unit satisfies our needs for now, but alternatives will be considered in further experiments.

3.2 Feature transformation

To increase performance, the feature space D can be translated to a different feature space D*, which generally has a smaller size. Two basic operations in feature transformation are deletion and merger.

Features that have no discriminative power may be considered for deletion. If words are used to define the feature space, it is also intuitively clear that a number of them do not contribute to a meaningful text similarity. Very common words such as articles and pronouns occur in any text, regardless of its contents, and can therefore be ignored. Other words occur only so rarely that their burden on storage size or retrieval time is too high to justify their presence (about 2/3 of the different words in our corpus occurred only once or twice in 7500 reports). Words can be deleted by using thresholds on word frequency, or on a computed 'discrimination value' for that word. The discrimination value for a word is defined as the difference in average text similarity (taken over a sample of text pairs) excluding and including that word. If the average text similarity remains about the same upon removal of

a given word, the discriminative power of that word is low. The average text similarity decreases if a very common word is deleted. Words that have a positive discrimination value are important: they help to distinguish between texts and therefore lower the average text similarity (Salton and McGill 1983). Deletion of features clearly shrinks the dimension space. It increases performance because the 'noise level' is reduced when irrelevant words no longer contribute to the similarity score.

Features that are highly correlated may be merged. Again, the dimension space becomes smaller when two features are merged into one. Performance increases because connected features that originally failed to contribute to the similarity score, do so after merger. For words, this happens when they express the same meaning, but have a different representation. This is the case for full synonyms, but also for words that have different inflections (caused by syntactical context), words that have different – correct or erroneous – spellings, and word concatenations. They can be transformed via a word list, a rule base or word pair matching (cf. Findler and Van Leeuwen 1979). The last two methods cannot be used to connect full synonyms.

Scaling 3.3

The importance of a feature in a certain vector can be stressed by assigning a higher or lower weight factor to that dimension in that vector. This will result in a more faithful representation of the sample, and therefore better retrieval results. In text matching the weight factor for a certain word in a given text may depend on the number of occurrences of the word in that text, the overall word frequency, the proportion of samples in which that word occurs, and/or the discrimination value (see description above) of that word (Sparck Jones 1973). Refined weight factor calculation demands that a word base is maintained. However, simpler weight factors may prove to give satisfactory results as well. The simplest of these is binary weighting: if a word occurs in a text, its weight is 1, otherwise it is 0. Other collection independent weight factors are word frequency within the text, or even word length (function words are often short words, content words are often longer words).

Performance of searching 4.

The outcome of a search action is usually given with two measures: precision and recall (Salton and McGill 1983). Precision is the proportion of relevant items in the total set of items that was retrieved. Recall gives of all relevant items the proportion that was retrieved. If all relevant items and nothing but relevant items are retrieved, both precision and recall are 1. The problem with measuring recall is, that the whole archive should be scrutinised to find items that were unjustly not

retrieved. The problem with both measures is, that 'relevance' is not always a binary state: items can be 'a bit relevant'. The set of relevant items cannot be strictly defined so recall and precision are measured less precisely (Salton et al. 1994).

In classification, recall is not that important as long as precision is good. If the (k-) nearest neighbors of a given sample give a suitable classification, it does not matter how many other samples with equally suitable classifications were rejected.

Performance varies with several factors:

corpus size. With a larger corpus, the chance increases of finding good neighbors for any given sample. A larger corpus however asks for larger storage capacities and probably longer retrieval time. And even our 5000 report archive proved inadequate to correctly classify about 3.5% of new pathology reports. Growth of the corpus could therefore better follow a careful strategy of adding rare cases and ignoring routine cases.

Corpus organisation. Items can be retrieved faster and possibly more accurately if the corpus is appropriately structured. Several methods can be used: clustering, inverse tabelling and using reference samples.

In clustering, the data is stored hierarchically in two or more layers. For retrieval, all highest level clusters are matched and the best one is entered, after which the second highest levels are considered – and so on (c.f. Young and Calvert 1974). The performance may respond to clustering in either way: the search algorithm might overlook relevant items, so that recall drops, but may also produce less false hits, which improves precision.

With inverse tabelling, a subset of features is used as the key to a table that gives the sample identificators (e.g. report numbers) of those samples that contain that feature. Upon searching neighbors for a new sample, those corpus samples that have many features in common with the new sample can be identified from the table. They would be the first candidates for an actual similarity score computation.

The 'reference sample' method is described by Sethi (1981): a number of reference samples is chosen or created. For every corpus sample, the distances to the reference samples are computed and stored in a table. For classification of a new sample, those corpus samples are first considered that have about the same distance to the reference point as the new sample.

The organisation methods also influence the other performance: *process performance*. Process performance includes the speed with which a sample is classified, and the requirements on storage capacity and memory size. In our application, the system should give a classification quickly after a text is entered or else it will lose the user's appreciation. The retrieval of the five nearest neighbors of a report out of a 5000-report archive took about two minutes on a standard personal computer (modest 1995 standards: a 486DX2-33, 8 MB RAM). Such a search action involved comparing the report with all 5000 (unordered) archive reports. This search time

already decreased to 3 to 10 seconds when we adopted an hierarchical archive organisation, lateron. Process performance is not our prime interest, and in the following experiments we concentrate on the functional performance.

Practice 5.

The pathology reports that are considered here, are histology reports – diagnoses of tissue examination. They contain paragraphs on clinical details of the patient (usually one or two lines), a short description of the source material, sometimes the questions that the clinician wants answers to, the macroscopical and the microscopical observations and the conclusions. The length of a report is typically about 150 words, but this ranges between ten words and a thousand words or more.

For successful classification of histology reports with the (k-) Nearest Neighbor rule, three assumptions should be met:

- ◊ *a given case is not unique because in the past, similar cases were diagnosed.* The scala of cases in pathology runs from very common/routine cases to very rare cases. The chances of finding a similar, rare case depends on the size of the archive that is searched.
- ◊ *identical cases are given the same classification codes.* Although this is the philosophy of the coding system, the formal language still has enough degrees of freedom to give different correct/valid classifications to a single case.
- ◊ *for the description of similar cases in natural language, similar phrases are used.* If textually similar reports describe similar cases, the same formal classification can be applied to textually similar reports.

A pilot experiment and three simulations were run to test these assumptions.

Pilot experiment 6.

Four variants of an algorithm were used to search nearest neighbors for five 'new' reports. An expert scored the report pairs, that were thus formed, on semantical similarity. These were used to identify the most successful algorithm and to estimate the usefulness of the generated classifications.

Material: an archive of 1180 histology reports was used: the entire production of January 1994 of the pathology laboratory of the university hospital in Maastricht. The archive was unclustered. For nearest neighbor identification of a given report, similarity scores were computed for each of the archive reports.

Five reports were randomly drawn under the restrictions that
◊ the report describes only one type/slide of material, and
◊ the five reports differed sufficiently from each other.
Those reports described:
I) subacute dermatitis (skin tissue),
II) cystous atrofia of endometrium
III) stomach corpus biopsies with superficial gastritis
IV) a varicous vene
V) sigmoid biopsies with chronic active inflammation

For these five reports, the nearest neighbors were searched with four algorithms:
alg 1: all words contribute to the similarity score
alg 2: all words contribute to the similarity score except very frequent words (we defined this as occurring in more than 20% of the reports).
alg 3: word variants were uniformised. For this, a manually constructed substitution table was used: 2493 words were replaced so that the archive contained 5825 different words afterwards (7962 before replacement).
alg 4: word variants were uniformised and very frequent words were ignored.
In all algorithms, binary weight factors were used.

Procedure: Per algorithm, five nearest neighbors were searched for the five 'new' reports. Another five archive reports were retrieved with keyword search and added to the set of nearest neighbors. Because some archive reports appeared in the top-5 nearest neighbors of two or more algorithms, the number of different retrieved archive reports per 'new' report was not 25, but less. These reports were randomly ordered, and given to the expert for scoring. Scores were given on three criteria:
◊ is the material in both reports the same,
◊ *is the tissue obtained in the same way*
◊ is the condition the same in both reports?
These criteria correspond with the main three aspects on which material is usually filed in pathology. Scores were given on an 11 point scale, from 0 to 10 (from totally different or contradictory to completely identical).

6.1 Results

The averages of the expert ratings are given in table 1. The results are weighted to the ranking of the report pair: the nearest neighbor is weighted with factor 5, the second nearest neighbor with factor 4 etc.

6.2 Discussion

1. relative retrieval performance: *what algorithm produces the best results?* Table 1 shows that algorithm 1 does not always retrieve optimum reports. For trial reports II and V, there are other algoritms that are able to find more relevant items in the

archive. If algorithm 1 does find good archive reports, for trial reports I, III and especially V, the results are not degraded upon additional operations (i.e. for alg. 2 to 4). The results suggest that elimination of the very frequent words is necessary for good retrieval, and using uniform words may give an additional improvement. With the limited data, trying to statistically prove these observations is not sensible.

2. absolute retrieval performance: *are the results from the best algorithm sufficiently good?* Only for trial report IV the hit rate was 100%. This trial report described a fairly routine case (a varicous vene). For trial report II, the expert rated the suggestions to be barely sufficient; for all other trial reports, the ratings indicated good performance. If the SNOMED codings are considered, it was found that 15 out of the 17 SNOMED terms that had been assigned to the trial reports earlier, appeared as first suggestions from alg. 2 and 4 – the other two SNOMED terms appeared as the first alternative.

With this experimental design it is difficult to judge imperfect performance. Either the archive does not contain better items, or it does, but they were not retrieved. In the first case, low ratings do not imply that the algorithm is inadequate – better performance can then be expected upon expanding the archive. Since the expert cannot be expected to rate every archive report against the trial report, figures for unretrieved relevant items remain unknown.

From the pilot experiment it is concluded that archive searching seems a promising method for classifying free texts.

TABLE 1: weighted averages of expert ratings (max, sd) for the four algorithms

	trial reports				
	I	II	III	IV	V
algorithm 1 (basic retrieval)	8.84 (10, 2.50)	2.07 (10, 3.39)	7.87 (8.3, 2.43)	10 (10, -)	6.98 (9.3, 2.98)
algorithm 2 (without frequent words)	8.73 (10, 2.77)	6.18 (10, 4.61)	7.84 (9.3, 2.48)	10 (10, -)	7.80 (9.3, 2.80)
algorithm 3 (uniformised)	8.69 (10, 2.82)	4.51 (7.7, 4.48)	7.51 (8.7, 2.90)	10 (10, -)	6.91 (9.3, 2.99)
algorithm 4 (uniform, without frequent words)	8.80 (10, 2.59)	5.76 (10, 4.85)	7.71 (8.3, 2.48)	10 (10, -)	8.22 (10, 2.60)

7. Simulation experiment

Material: since the pilot experiment data indicated that the archive was not large enough to cope with moderately rare cases, we enlarged our corpus to 5000 reports. These were all the histology reports that were produced during (about) the first term of 1994 at the pathology department, Academic Hospital Maastricht.

Evaluation criterion: The evaluation method of the pilot experiment is not suitable to evaluate large numbers of trial reports. Therefore an evaluation criterion was constructed that allowed simulation experiments to be conducted. All the archive reports have been annotated with SNOMED terms[1], so that for a given report pair the semantical similarity can be estimated through shared terms in the SNOMED lines. This can be represented in a figure with the common formula (c.f. Findler and Van Leeuwen 1979):

$$S_{1,2} = \sum f_{1i} f_{2i} \; / \; (\sum f_{1i}^{\,2} \; \sum f_{2i}^{\,2})^{1/2} \qquad (2)$$

Or, the SNOMED score S for report pair 1 and 2 equals the multiplied frequencies of shared SNOMED terms i, normalised between 0 and 1.

Since our expert already scored 77 report pairs, the calculated SNOMED score for these report pairs could be put against the expert's ratings. This results in figure 1: a scatter plot of expert ratings against SNOMED ratings. The correlation coefficient between the two ratings is .76 (p=.000). Figure 1 indicates a vulnerability of the SNOMED rating in the middle region (between .32 and .56). In this region, the SNOMED rating gives a poor prediction of the expert's rating. Cause of the poor prediction here, lies probably in the degrees of freedom the SNOMED coding system allows.

Simulation method: the retrieval method that is evaluated is not – or only marginally in some model variants – dependent on archive characteristics. Leave-one-out simulations can therefore be used: one report is lifted from the archive and considered to be the trial report. In retrieval, the entire archive minus the trial report is then searched. In the simulations that are described in this article, the leave-one-out simulation is repeated for 500 consecutive archive reports from the total of 5000 reports.

---

1) a typical report is annotated with three or four SNOMED terms (for the 5000-report archive: median=4, mean=4.22 (sd=2.06), range=[1..25]). They describe topology, morphology and source of the material, and sometimes also aetiology and function. A SNOMED term consists of a 6-character formal string, e.g. 'M43000' or 'TYY990'. In the archive, a total of 983 different terms occurred - 463 of these were encountered only once or twice.

Simulation 1: basic retrieval 7.1

For this simulation, binary weights were applied. All words contributed to the similarity score, but numerals were removed. For each of the 500 repetitions, the five nearest neighbors were retrieved and their SNOMED terms were compared to the SNOMED terms of the 'seed' report. For the five SNOMED ratings that resulted from these comparisons, the weighted average and the maximum were determined. The weighted average gives an estimation of the overall quality of the top-5 retrieved reports. The maximum SNOMED rating is relevant when the method is used to retrieve candidates for further processing.

Retrieval performance is bounded by the contents of the archive. That is why the SNOMED ratings for a seed report should be related to the highest possible rating that is achievable given the archive. Figures 2 and 3 show the distributions of the averaged and maximum SNOMED ratings for the retrieved reports, and their respective highest possible ratings.

Results: the figure shows that for 79% of the seed reports a successful archive report (SNOMED rating >.56) was found in the top 5. The top 5 contained no successful archive report for 7% of the seed reports – the remaining 14% is located in the middle region for which no conclusions may be given. For 52% of the seed reports, the averaged SNOMED rating exceeded the .56 boundary. For 18% the top 5 was inadequate on average; for the other 30% no clear judgment can be given.

FIGURE 1: expert rating vs. SNOMED rating

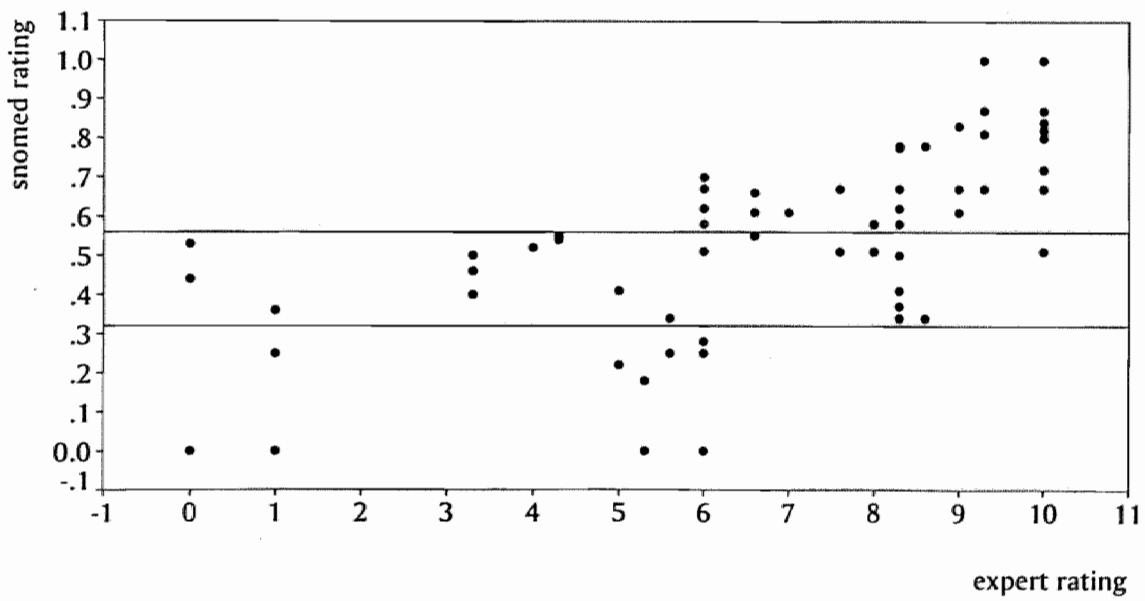


FIGURE 2: averaged SNOMED ratings: distribution plot – binary weight factors

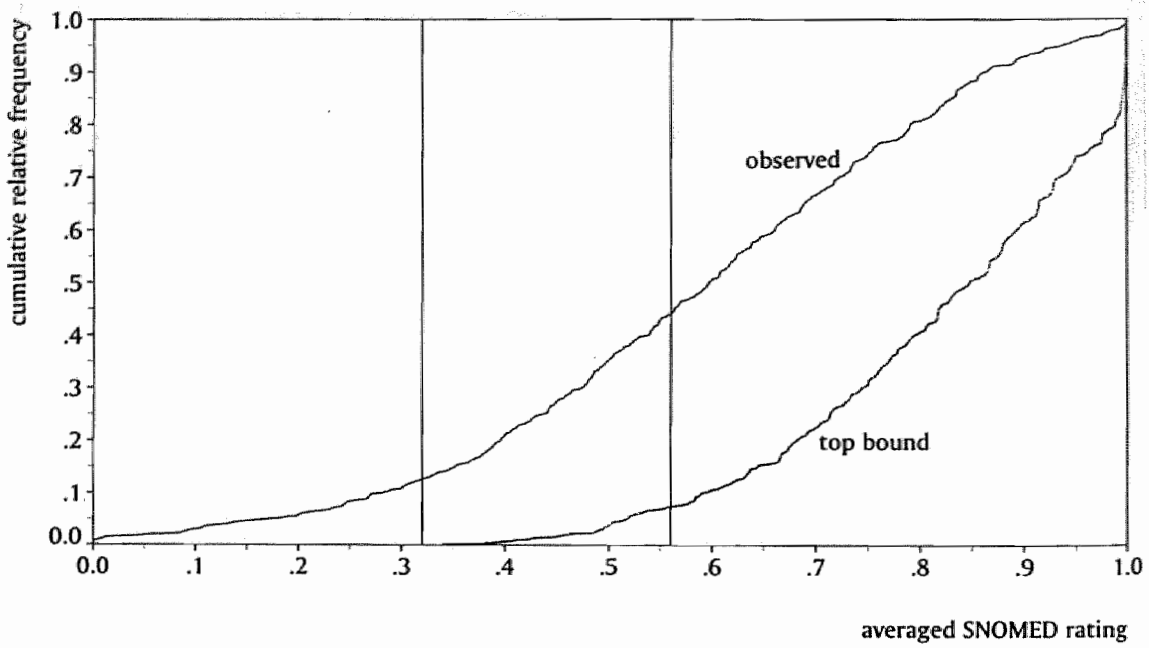
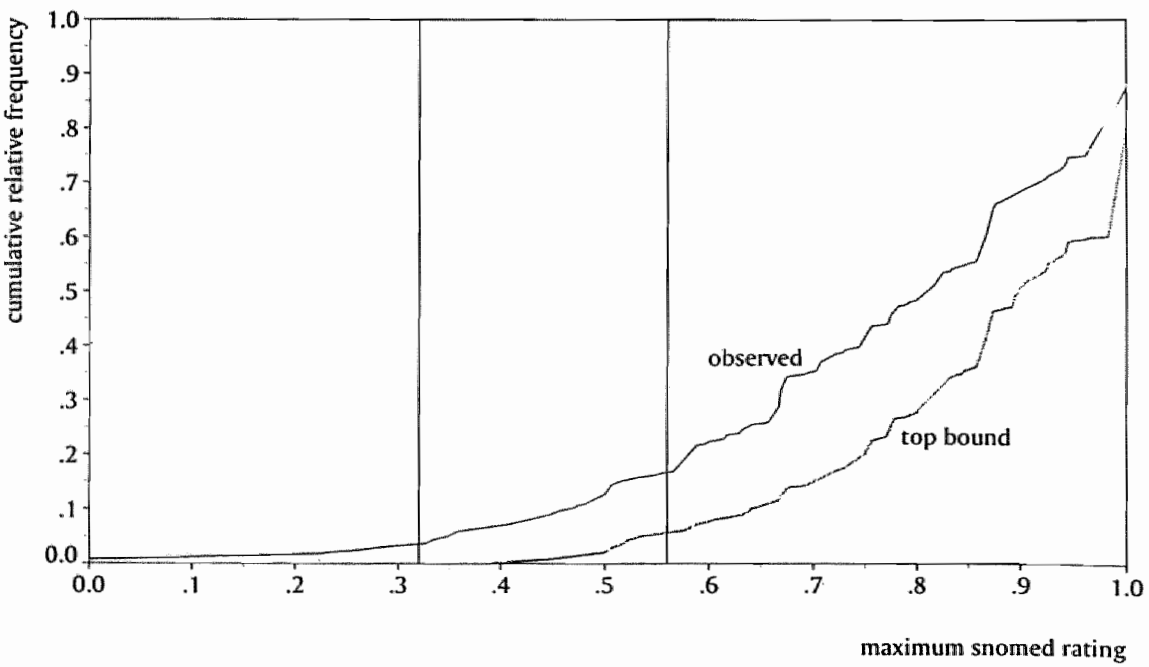


FIGURE 3: maximum SNOMED ratings: distribution plot – binary weight factors



The total performance of retrieval can be represented in a coverage percentage: that part of the top bound area in the figures that is covered by the actual retrieval, or the actual retrieval rating divided by the highest possible rating, integrated over the 500 repetitions. Thus the coverage percentage for the averaged SNOMED ratings is 65.2%, for the maximum SNOMED rating it is 82.9%.

Simulation 2: enhanced basic retrieval 7.2

Retrieval with binary weights is repeated, but after elimination of very frequent words – those words occurring in more than 20% of the archive reports.

Results: performance for this simulation was slightly but not significantly better than for basic retrieval (65.5% coverage for averaged SNOMED ratings, 84.1% for maximum SNOMED rating). Significances for this and the following simulations were determined with paired t-tests; $p<0.005$.

Another enhancement consisted of uniformising the word forms, such as verb conjugations and declensions of adverbs and nouns. For this, a manually constructed stop list was used.

Results: for this simulation, performance again was slightly but unsignificantly better than for basic retrieval (66.0% coverage for averaged SNOMED ratings, 83.6% for maximum SNOMED rating). No significant differences were found with retrieval without the very frequent words.

The third enhancement consisted of again uniformising the word forms, and then identifying and eliminating the very frequent words (>20% of the reports).

Results: no significant differences were found with any of the previous simulations. Coverage was 65.9% for averaged ratings, and 83.7% for maximum ratings.

Simulation 3: non-binary weights 7.3

The simulations with all words, all words without the very frequent ones, the uniformised words and the uniformised words without the very frequent words were repeated with the use of a non-binary weight per word. The weight for a certain word in a report is the frequency of that word in the report, multiplied with $\log (R / R_w)$ – R is the number of reports (in our case 5000), R_w is the number of reports in which word w occurs. These weight terms are analogous to those defined in (Salton et al. 1994). Over the 500 repetitions per simulation, weight terms are kept constant, but are recomputed after uniformation of the words.

The simulation with all words, using non-binary weight terms, resulted in significantly higher ratings than any of the simulations with binary weights. The characteristics are shown in figures 4 and 5: distribution plots of the SNOMED ratings. The averaged ratings covered 68.5% of the highest achievable ratings, the maximum ratings from the top-5 covered 87.1% of the highest achievable ratings.

FIGURE 4: averaged SNOMED ratings: distribution plot – non-binary (and binary) weight factors

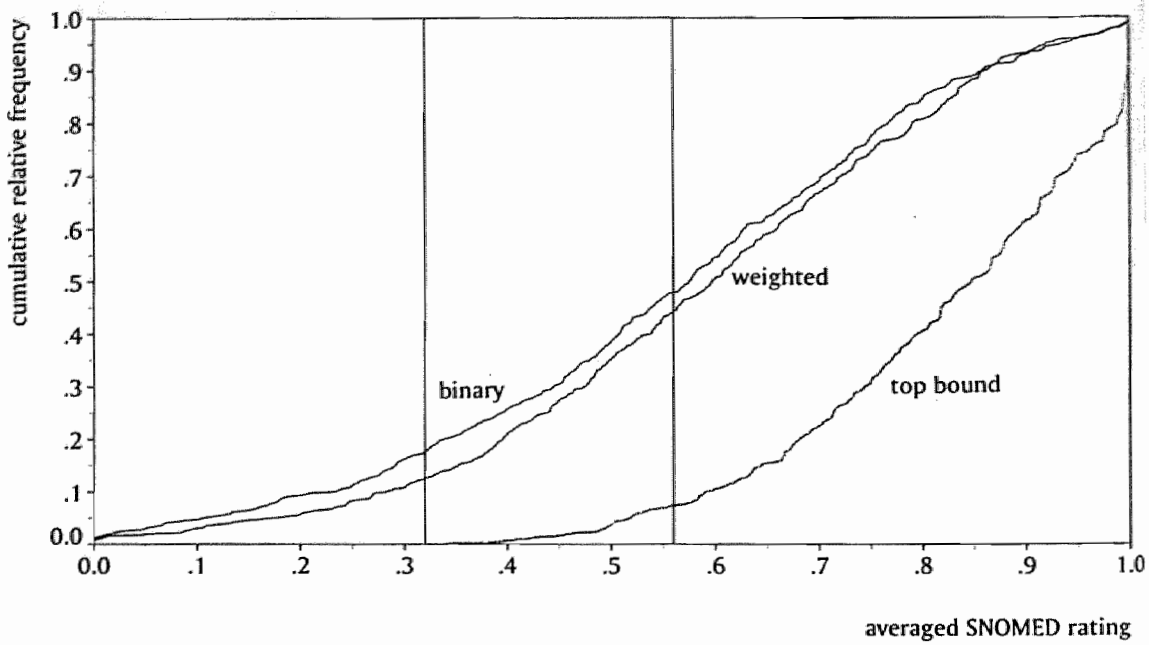
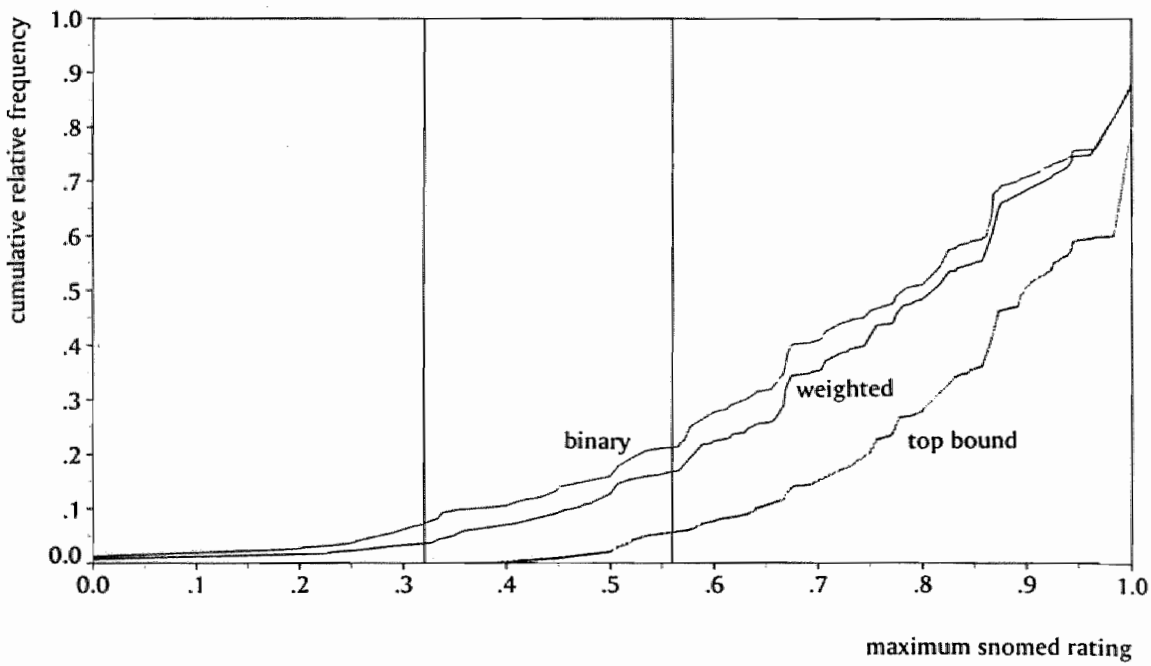


FIGURE 5: maximum SNOMED ratings: distribution plot – non-binary (and binary) weight factors



Results: the figures show that with the weighted model for 83% of the seed reports, a good archive report was retrieved within the top-5, for 3.5% the top-5 included no good items. For the remaining 13.5% no judgment can be given. For 56% of the seed reports, the averaged SNOMED rating exceeded the .56 boundary. For 13% the top 5 was inadequate on average; for the other 31% no clear judgment can be given.

None of the enhancement operations resulted in significantly better retrieval (paired t-test, $p<0.005$). Retrieval performance did not differ between any given pair of methods except for slightly higher averaged SNOMED ratings with the uniform-word model in comparison with the model without very frequent words ($p=0.004$). Each of the weighted models differed significantly with any of the binary models ($p<0.005$), for the averaged and for the maximum ratings.

7.4 Discussion

In general it can be said that the model gives satisfying results. For a large proportion of the experiment repetitions (83% in the best model), appropriate archive reports could be found amongst the five retrieved items.

Cases in the lower region of the curve, those with low SNOMED ratings, were indeed the reports where the archive did not contain any similar SNOMED line. An expert in pathology confirmed that these reports did describe unusual cases.

For the region between .32 and .56, further experiments must give clarity: the SNOMED rating criterion is too vulnerable in this region. Some report pairs that score in this region, will prove to be highly similar when scored by an expert. In a larger scale experiment, therefore, expert ratings will again be used.

The additional operations (viz. uniformising words, removing very frequent words) that were used, yielded no improvement in performance. Results were not poorer either, however, so these methods can be used for dimension reduction without loss of performance. This means less storage and less online computation, which can be advantageous, but does demand more preprocessing.

The use of weight factors gives an important gain over the binary model. The problem with these weight factors is, that a lexicon is needed. Such a lexicon must be maintained to cover updates in the archive. Weight factors could stable out after a period of time, so refreshing of the weight list is not a daily question, but should be considered only every once in a while. This would enable to accurately incorporate new trends in word use. Other weight models than the log of the inverse report frequency will be considered in further experiments.

The result of the nearest neighbor search is not suitable for unsupervised automatic classification in our practical setting. It does give good input for further processing steps, though. A polling scheme could decide upon the final outcome from several near neighbors. Natural language processing techniques could complement the nearest neighbor search, which forms the hybrid system with both statistical and knowledge-based elements that Croft (1993) advocates. In our practical application,

the user can perform postprocessing by accepting or rejecting the first suggested alternative, considering a limited number of other alternatives or otherwise altering the outcome.

8. Conclusions

A method is presented that classifies natural language texts through nearest neighbor search. The success of the method hinges on the degree in which a computed textual similarity predicts the semantic similarity of a text pair. Two experiments confirmed that the nearest neighbor method was able to retrieve relevant archive texts for a trial text. In the pilot experiment, an expert rated the retrieval results: these were good to excellent for four out of five trial reports, and fair for the fifth one. In the simulation experiments, it was shown that for most of the trial reports (83% for the best model), the five retrieved items contained an appropriate archive report. Weight factors that stressed the importance of less frequent words, were found to give a significant improvement over the model with binary weights. Elimination of very common words and uniformation of word variants did not change retrieval performance either way. These operations will therefore not improve the outcome, but may result in better system performance with equal functional results.

References

Cover T.M. and Hart P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory Vol IT-13 1967 pp 21 - 27

Croft W.: Knowledge-based and statistical approaches to text retrieval. IEEE Expert 1993, pp 8 - 12.

Damashek M.: Gauging similarity with n-grams: language-independent categorization of text. Science, Vol 267 1995 pp 843 - 848

Findler N.V. and Van Leeuwen J.: A family of similarity measures between two strings. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol PAMI-1 1979 pp 116 - 118

Fukunaga K.: Introduction to Statistical Pattern Recognition, 2nd edition. Academic Press, 1989.

Sethi I.K.: A fast algorithm for recognizing nearest neighbors. IEEE Transactions on Systems, Man and Cybernetics Vol SMC-11 1981, pp245-248.

Salton G. and McGill M.J.: Introduction to modern information retrieval. Addison Wesley, 1983.

Salton G., Allan J., and Buckley C.: Automatic structuring and retrieval of large text files. Comm. of the ACM Vol 37 1994 pp 97 - 108 (1994a)

Salton G., Allan J., Buckley C. and Singhal A.: Automatic analysis, theme generation, and summarization of machine-readable texts. Science Vol 264 1994 pp 1421 - 1426. (1994b)

Schalkoff R.J.: Pattern recognition: statistical, structural and neural approaches. Wiley New York, 1992.

Sparck Jones K.: Index Term Weighting. Information Storage and Retrieval V9, 1973, pp919-633.

Van Rijsbergen C.J.: Information Retrieval – second edition. Butterworths, London, 1979.

Young T.Y. and Calvert T.W.: Classification, estimation and pattern recognition. Elsevier, New York, 1974.

CHAPTER · 4

AUTOMATIC CLASSIFICATION: COMPARISON OF DIFFERENT MODELS

Part 1: Performance for various types of reports.
Part 2: Influence of archive collections.
Part 3: Word based model vs. n-gram model

Part 1: Published as:
L.M. de Bruijn, A. Hasman, and J.W. Arends: Automatic SNOMED classification — a corpus-based method. Computer Methods and Programs in Biomedicine 1997 Vol 54 pp 115-122

Part 2: Submitted for publication

Part 3: Submitted for publication

4 AUTOMATIC CLASSIFICATION: COMPARISON OF DIFFERENT MODELS

Part 1: Performance for various types of reports.

1. Introduction

A written report drawn up in natural language is the preferred means of communicating the results of an examination between a diagnostician and a physician: free text narrative is powerful yet comfortable for both the reporter and the recipient. The formal nature of the computer, on the other hand, requires that communication with clinical databases takes place in a formal language. Medical coding systems such as ICD (ICD 1992), SNOMED (Gantner et al. 1979) and Read Codes (Read 1990) have been developed and implemented in order to guarantee accurate storage and retrieval of cases. They form the key to open doors to quality control, patient monitoring and informational support in clinical and epidemiological research.

Coding of diagnoses is a difficult and time consuming task for the reporter, and the archiving organisation has to cope with errors and subjective variation in coding. So ever since the introduction of formal coding systems, efforts have been made trying to automatically connect the natural language description of a case and its appropriate formal coding terms (Altman and Oliver 1994, Wingert 1985, Sager et al. 1982, 1987, 1992). A number of disadvantages of existing methods led us to adopt another approach, which is presented in this article.

2. Method

Automatic text classification can be seen as a specific case of pattern classification, one in which the patterns are textually oriented. There is a choice between two approaches: a structural and a statistical one (Schalkoff 1992). Parsing a text into syntactical structures and then applying knowledge rules to come to a classification is such a structural approach and one that is generally referred to as Natural Language Processing (NLP). In the Linguistic String Project (Sager et al. 1982, 1987, 1992), English pathology texts were processed into NLP algorithms. Since they require to be moulded to the domain on which they are applied, very accurate results can be expected *within the domain*. It also means that every extension of the domain, or change within it, demands (often manual) adaptation of the set of rules that make up the NLP system. These NLP-based systems are language dependent.

The Nearest Neighbor method that we use in our research falls in the second category: statistical text classification. Upon entry of a new report, the archive of reports is searched for similar reports. If the archive reports that were found are indeed not only similar in appearance, but also in meaning, then the classification codes of the archive report are candidates for classifying the new report.

This approach hinges on two conditions:

- ◊ a similarity measure can be constructed such that a high value guarantees that the reports describe highly similar cases. This condition implies that the initial *classification* problem has now become a *search* problem.
- ◊ an archive is available that is annotated with classification codes, and that contains a sufficiently diverse range of reports so that many different new reports can be handled. Note that most medical departments already keep an annotated archive in electronic form.

The method itself is independent of the domain, it can be applied in any language, any classification system and any medical or non medical discipline. The material we used for our research was written Dutch pathology (histology) discourse, and coded in SNOMED terms. For the sake of universality, the procedures in this setup are kept as general as possible: we did not enhance performance by including language dependent transformations, by tuning weight factors through (domain dependent) relevance, or by profiting from direct relations that can be made between words in the texts and descriptive terms in SNOMED.

Theory: definition of the similarity measure 3.

The collection of archive reports can be thought of as a vector space, in which each report represents a vector in that space. The similarity between reports can be seen as the distance between two points in the space, or their closeness.

Vector space definition 3.1

The vector space is a multi-dimensional space, in which every dimension describes a feature that may be absent or present (or present to a certain extent) in the report. Since the only thing we know about the report is its appearance, the features are based on the words or terms that are used in the report's descriptions. The choice of a vector model implies that there are no temporal connections between dimensions. In the practice of report analysis this means that context is removed: the words in the report are disconnected from their phrase structure. This presumes that part of a text's content lies in the words per se.

The *choice of axes* we made is all the *words* in the reports. Other options would be word n-grams, or sub-word units such as syllables; *words* however is the most

natural choice here. In the collection of 7,500 histology reports that was used in this study, about 20,000 different words were encountered.

After the dimensions are determined, the *vectors are scaled*. With this operation, weight factors are assigned to the features so that important words have a higher influence in the similarity computation than unimportant words. The weight factors could be manually determined and compiled into a weight factor list, but a more objective way is to compute them through statistical observations on the total archive collection (c.f. Sparck Jones 1973). The weight factor computation that we used in our study is:

$$w_{i,j} = f(i,j) * \log (n / n_i);$$

that is, the weight factor $w_{i,j}$ for word i in report j is the number of occurrences f (i,j) of word i in report j, multiplied with the log of the total number of reports in the archive (n) divided by the number of reports in the archive that contain word i (n_i). This weight factor favours words that occur often in a certain report but rarely in the rest of the collection. Common words such as *'the', 'and', 'in', 'are'*, get low weight factors because of the high value for n_i.

The similarity measure that defines the nearest neighbor(s) of a given or new report is derived from the following computation (a vector inner product):

$$T(r1, r2) = \sum_i w_{i,r1} * w_{i,r2} \quad ;$$

which is then normalised to a figure between 0 and 1 with

$$T_n(r1, r2) = T(r1, r2) / \sqrt{(T(r1, r1) * T(r2, r2))} \quad .$$

If T_n (r1, r2) = 0, then two texts are very dissimilar, if T_n (r1, r2) = 1 then the vocabularies in the two texts are identical – which makes them very similar.

3.2 Vector space mapping

To improve performance, the feature space can be *mapped* onto a different feature space, which generally has a smaller dimensionality. The two basic operations in feature transformation are *deletion* and *merger*.

Features (words) that have no discriminative power may be considered for deletion. In context-free comparison, the function words such as articles and pronouns do not contribute to a meaningful text similarity, and can therefore be ignored. Other words occur only so rarely that their burden on storage size or retrieval time is too high to justify their presence. In our 7500-report corpus, about 13.000 of the 20.000 different words occurred only once or twice. Deletion of features shrinks the dimension space, which improves technical performance. Storage needs decrease: the ten most frequent words in our collection were

together responsable for more than 22% of the word occurrences. Functional performance improves because the 'noise level' reduces when irrelevant words no longer contribute to the similarity score.

Features that are highly correlated may be considered for merger. Again, the dimension space becomes smaller when two features are merged into one. Words may express the same meaning but have a different representation. This is the case for full synonyms, but also for words with different inflections (varying with syntactical context), or due to spelling preferences, typing errors and word concatenations. They can be transformed via a word list, a rule base or word pair matching (cf. Findler and Van Leeuwen 1979). The last two methods cannot be used to connect full synonyms. Functional performance increases because connected features that originally failed to contribute to the similarity score, do so after merger.

After transformation, the weight factors should be recomputed.

Experimentation 4.

Hypothesis 4.1

'The nearest neighbor method can correctly classify natural language reports through textually similar archive reports'.

Material 4.2

A collection of 7500 histology reports was obtained from two laboratories: the pathology laboratory of the University Hospital in Maastricht (AZM) extracted 5000 reports from their archive, and the laboratory of Elkerliek hospital in Helmond provided us with 2500 reports. For both locations this was the whole (unfiltered) production of about the first trimester of 1994. A typical report is composed of 80 to 130 words in Dutch medical discourse; this length ranges between 10 and 1300 words. A report contains paragraphs on the source of the material, some clinical data on the patient, the original query of the requesting physician, macroscopical and microscopical observations and the concluding diagnosis. All the archive reports were beforehand transformed into word lists and annotated with weight terms, as was described in the Theory section. Words were not transformed in order to keep the results as pure as possible, but numeral values in texts were kept out of the analyses.

The archive reports were separated from their actually assigned SNOMED codes, so that these codes could form the basis of an independent SNOMED evaluation (see below). The SNOMED codes together describe the organ, the site and the diagnosis of the case.

4.3 Trial reports

The simulation was repeated 1000 times: report numbers 501 to 1000 from the corpori of each of the two laboratories served consecutively as trial report. These portions were verified to be representative samples of the collections.

4.4 Simulation Method

A simulation was run with the 'leave one out' technique. This means that one report is lifted from the archive and serves as a trial report. This report is classified by searching nearest neighbors from the remainder of the archive. For the next trial, the former test report participates again in the archive.

Per trial, the top five most similar archive reports were retrieved by comparing the trial report with each of the reports in the archive collection. The simulation was performed using the whole text of the trial report and the archive report with untransformed words, repeated using uniform words, and once more performed using only the 'conclusions' paragraph of all the reports, with uniform words. The criterion for retrieval was the similarity computation as was described in the *theory* section, a measure that will be referred to in the remainder of this article as 'text similarity'.

4.5 Evaluation Method

With the desired number of repetitions, it was not achievable to collect expert ratings for every pair of reports in the experiment. Therefore we used data from a pilot experiment to construct an alternative evaluation measure, based on the actually assigned SNOMED codes [12]. A comparison by SNOMED terms between the two classification lines (from the pair of reports) yields a numerical evalua-tion score, which was assigned to one of the following three discrete regions:

- ◊ a lower region in which none or only a few of the SNOMED terms correspond between reports: the retrieved report is most probably irrelevant,
- ◊ a middle region in which the result is only partly relevant, dubious or uncertain,
- ◊ and a higher region in which most of the terms correspond, so that the report is most probably relevant.

The result will be referred to as 'SNOMED-rating'. For each trial the first nearest neighbor was assessed, as well as the next four nearest neighbors. For some of the trials, the archive did not contain a relevant solution, therefore the results were related to the maximum possible SNOMED rating - given the archive collection (c.f. the second condition that was stated in the *method* section).

Thus, the measure of evaluation was a *precision* measure only: *Recall* – analogous to the medical term 'sensitivity' – was not under consideration because it is not necessary to retrieve *all* relevant reports when one is enough. Moreover, the condition 'a high text similarity should guarantee that the reports describe highly similar cases' from the *method* section needs logically not be reversible.

Apparatus 4.6

For the simulations, we used a standard PC configuration (modest 1995 standards: 486DX33, 8MB). The data required about 9 megabyte disc space (about 2.5 megabyte compressed). In this configuration, the simulation per trial took at most 2 minutes. In these simulations, a trial consisted of calculating the text similarity with each of the collection reports. With a cluster-structured archive collection and a Pentium-100 configuration, retrieval time for a trial was reduced to about a second. Programs were written in Borland Turbo Pascal 6.0.

Results 5.

Table 1 shows the simulation results for classification with the nearest neighbor method on whole texts. In 618 of the 1000 trials, a relevant classification was given as a first suggestion. In 48 cases, the trial report apparently was a 'rare' case and the archive contained no relevant other reports at all. The relative retrieval performance was therefore 618 / 952 = 64.9%. In 830 trials, a relevant classification could be found within the first five suggestions, which gives a relative performance of 87.2%. In 3.6% of the trials, none of the first five suggestions contained a relevant classification. The remaining 13.4% is found in the 'uncertain' category.

Figure 1 displays the extent to which the first up to the first five nearest neighbors contain relevant classifications for the 1000 trials. A 'ceiling' is indicated at 952 trials: this is the maximum of cases that could be classified given the collection.

Retrieval through the transformed word list – where prefixes and suffixes are removed, spelling differences are equalised and word concatenations are disconnected – gives the results that are given in table 2. The differences with the figures in table 1 are not significant ($p > 0.1$; paired t-test).

If the matching between reports is done on the *conclusions* paragraph only, a relevant classification is given as a first suggestion in 596 trials; in 806 trials a

TABLE 1: basic retrieval performance; 1000 trials

number of nearest neighbors included	number of trials (proportion of maximum) rated relevant	rated uncertain/ partly relevant	rated not relevant
1	618 (.65)	239	143
2	744 (.78)	182	74
3	792 (.83)	154	54
4	814 (.86)	141	45
5	830 (.87)	134	36
maximum	952	48	

FIGURE 1: retrieval performance for n nearest neighbors

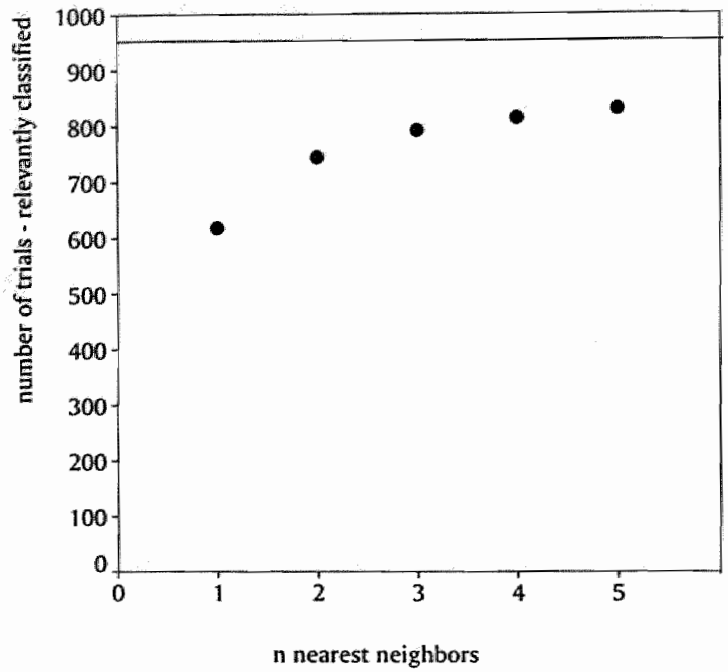


TABLE 2: retrieval performance with transformed list; 1000 trials

number of nearest neighbors included	number of trials (proportion of maximum) rated relevant	rated uncertain/ partly relevant	rated not relevant
1	626 (.66)	236	138
5	844 (.89)	122	34
maximum	952	48	

TABLE 3: retrieval performance with 'conclusions' only; 1000 trials

number of nearest neighbors included	number of trials (proportion of maximum) rated relevant	rated uncertain/ partly relevant	rated not relevant
1	592 (.62)	214	187
5	801 (.84)	140	52
maximum	952	48	

relevant suggestion could be found within the first five alternatives. In 7 trial reports, the *conclusions* paragraph contained no text, and classification is thus impossible. Table 3 summarises the results of this simulation. These results differ significantly with those in table 2 – full-text retrieval with transformed words – ($p<0.05$; paired t-test), but no significant difference was found with basic retrieval (table 1; $p>0.05$; paired t-test).

Discussion 6.

Results of basic retrieval 6.1

In 618 of the 1000 trials with a pooled archive, a relevant solution was given as a first suggestion for classification. Another 212 trials could be classified if the next four alternatives were also considered. Figure 1 shows that the 'ceiling' – in this case 952 reports that were classifiable – is rapidly approached with the first three neighbors, after this, retrieval performance levels out. For only 38 cases, the fourth or fifth suggestion was relevant after three partly- or non-relevant suggestions.

If the method is used with the first five nearest neighbors, it could cope with 830 out of 952 (87.2%) classifiable trials, or a total of 83.0%. Sager et al. (1992) reported correct classification in 107 out of 150 cases (71.3%) with their syntax-oriented method. In their experiment 25 reports did not reach the coding stage, so out of the 125 reports that were coded, performance was 85.5%. The methods of both experiments differ, as does the domain (head cancer reports in their study, a large range of different histology reports in our study) and the reporting language (English vs. Dutch), therefore we draw no further conclusions from this comparison.

Results for various types of reports 6.2

Multiple reports are difficult for retrieval – these are cases in which sections from two or more different locations are examined; typical combinations are stomach + colon, stomach + oesophagus, appendix + gall bladder, and endometrium + endocervix. A multiple report may succesfully retrieve another multiple archive report. However, the retrieval method discards context information, so that chances are that the final diagnoses are swapped between the two halves (or several segments). For instance, a report that described superficial gastritis with helicobacter infection in the stomach's pyloric region and helicobacter infection without inflammation in the stomach corpus showed prone to retrieve other 'stomach reports' with varying degrees of colonization and inflammation. In another instance, a report describing fibrosis in myocardium *and* endocardium received suggestions for the 'myocardial fibrosis' only – no other endocardium reports were present in the collection. In order to cope with these cases, the retrieval method should be enhanced with a routine that separates paragraphs or

sentences on the different slides or a routine that incorporates the proximity of words in the computation, which would make the method more complex.

A case with a certain disease can retrieve a nearest neighbor in which tissue is suspected for that disease but does not show enough evidence to diagnose it as such. The method naturally retrieves such a report; this would explain why for a number of trials a highly similar text was rated less relevant. Inclusion of the second, third, fourth or fifth suggestion in evaluation reduces the risk of such cases leading the classification algorithm astray. In the simulations, the retrieval method proves very efficient to deliver a low number of highly potential candidates, and a post-processing mechanism might select the best of those candidates. Post-processing could be performed with enhanced text-analysis methods, or it could be a task of the user, who selects the most appropriate classification from the set of three or five options. Such a hybrid system could well be the optimal solution for the classification problem (c.f. Croft 1993).

For some of the trials, useful classifications were given, but the SNOMED rating was incapable of showing this and placed the trial in the 'uncertain' column. For instance, 'skin * lipoma' is essentially a good classification for a case that was really coded as 'finger * lipoma'. On the other hand, 'skin * inflammation' is really different from 'skin * verruca'. Both combinations would result in the same SNOMED rating.

The method that is presented here, can be put into use as a module that captures a diagnostic report from the word processor, and suggests suitable classification phrases for the case.

6.3 Retrieval after word transformation

Word transformation results in a slightly better success rate, although this difference was not found to be statistically significant. Words can be transformed by using stemming algorithms and spelling check functions, but also by using a word transformation list. Since a weight factor list has to be consulted in retrieval anyway, this latter option does not burden the technical performance.

6.4 Retrieval with 'conclusion' text only

The simulations using the *conclusions* paragraphs only, show that the results are poorer than with retrieval on the whole text with transformed words. The reduction in accuracy can be accepted if higher priority is given to reducing data storage and retrieval time. The lower results are due to the shortness of the *conclusion* lines – which are often only a few words long: the 'signal' level drops if different phrases are used to state the same conclusion, and the 'noise' level rises when less important words unfortunately co-occur. But note that in practice the drop from a 88.7% success rate to 84.1% will be experienced as an increase of 40% in modifying-actions: necessity for debugging the classifications raises from 11.3% to 15.9% of the cases.

7. Conclusions

In short, the nearest neighbor method is a versatile approach for classifying texts. The experiment results indicate that the method is suitable for its purpose: suggesting potentially good classifications to the reporting diagnostician. The method itself is independent of the language or contents of the texts: these aspects are determined by the contents of the archive. Qualities such as accuracy, speed and flexibility can therefore be varied to a great extent by manipulating the archive collection.

Further exploration of the approach is expected to eliminate the need of coding classification for individual reporters and enhance uniform, objective, consistent and full codes for computerised retrieval in the near future.

References

Altman R.B. and Oliver D.E.: Extraction of SNOMED Concepts from Medical Record Texts. Proceedings of JAMIA Eighteenth Annual Symposium on Computer Applications in Medical Care. November 5-9, 1994 Washington D.C. pg. 179.

Croft W.B.: Knowledge based and statistical approaches to text retrieval. IEEE Expert 1993 pp 8 - 11.

De Bruijn L.M., Hasman A., Verheijen E., Van Nes F.L., and Arends J.W.: Classification of diagnoses that are described in natural language. Int. J. of Technology Management (in print) 1997.

Findler N.V. and Van Leeuwen J.: A family of similarity measures between two strings. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol PAMI-1 1979 pp 116 - 118

Gantner G.E., Côté R., and Beckett R.S.: Systematized Nomenclature of Medicine Coding Manual, College of Americal Pathologists, 1979.

International Statistical Classification of Diseases and Related Health Problems: ICD, tenth revision. World Health Organisation, Geneva 1992.

Read J.: Read Clinical Classification. BMJ 1990 pp 301-345

Sager N., Bross I.D.J., Story G., Bastedo P., Marsh E., and Shedd D.: Automatic encoding of clinical narrative. Comput. Biol. Med. Vol 12 issue 1, pp 43-56, 1982.

Sager N., Friedman C., and Lyman M.S.: Medical Language Processing: Computer Management of Narrative Data. Addison Wesley, Reading MA 1987.

Sager N., Lyman M., Nhan N.T., and Tick L.J.: Medical language processing: applications to patient data representation and automatic encoding. Methods Inf Med 1995 Vol 34 pp 140 - 146.

Schalkoff R.J.: Pattern recognition: Structural, Statistical and Neural Approaches. Wiley New York, 1992.

Sparck Jones K.: Index term weighting. Information storage and retrieval Vol 9 pp 619-633, 1973.

Wingert F.: Automated Indexing Based on SNOMED. Methods Inf Med 1985 24(1) pp 27-34.

4 AUTOMATIC CLASSIFICATION: COMPARISON OF DIFFERENT MODELS

Part 2: Influence of archive collections.

1. Introduction

In many places, medical narrative has survived the trends of formalisation that came with the information technology boom. Not without reason: it offers more freedom of nuances and detail than any formal description system, and it is the most natural form for communicating a diagnosis for both the reporting and the receiving person. Still, most diagnostic laboratories could not entirely escape the information managers, who did ask for a formal representation of the diagnosis for computer storage to accompany the natural language report. Pathologists, for instance, are confronted with the task of coding diagnoses into the formal language that their laboratory adopted, such as SNOMED, ICD or CMIT (Blois 1984).

The advantages of centralised computerised management of medical diagnoses are beyond dispute. Diagnoses can be retraced for even the most migrant patients, and access to computerised archives makes it possible to perform clinical and epidemiological studies that were impossible, unmanageable, or unreliable before.

The alliance between computerised storage/retrieval of documents and a formal coding language is strong, but not definitive. It may well be possible that natural language analysis methods will eventually evolve far enough to make formal coding no longer necessary or hide it from the user. The method that is presented in this paper does not go that far yet. It aims to process diagnosis texts, and suggest appropriate classifications or classification terms to the reporter. This means that the reporter is helped in a difficult task – i.e. making up a formal classification that is in form *and* in meaning correct, complete and elegant – without having to consult the coding-dictionary. The archiving organisation can expect to receive diagnostic codings that are possibly more dependable and probably less subject to personal preferences of the reporters. As it stands now, they have hardly any instruments to detect semantical inconsistencies in a classification.

2. Method

A 1982 article by Sager et al. (1982) describes the group's research on 'automatic encoding of clinical narrative'. The paper illustrates one way to attack the 'automatic coding' problem: by parsing the text into grammatical relations between syntactical units. This is known as the Natural Language Processing (NLP) method.

In spite of the advances in four decades of research, some drawbacks of the method have prevented widespread use of the techniques. NLP algorithms require to be moulded to the domain on which they are applied. This means that a NLP system is able to give very accurate results, *within the domain*, once the modelling is done. It also means that every extension of the domain, or change within it, demands (often manual) adaptation of the set of rules that make up the NLP system. Flexibility is therefore not one of the method's strong points.

An alternative classification method stems from research on Information Retrieval (IR) (Van Rijsbergen 1979, Salton and McGill 1983). In IR, the statistical properties of a text are used to extract keywords that are relevant to the meaning of the text. In contrast to NLP, a text is not regarded on its own, but against (the statistical properties of) an archive of texts. If a new text can be placed alongside an archive text that has exactly the same meaning, then classification is only a matter of borrowing the class code of the new text's neighbor. In the terminology of Pattern Recognition theory – to which IR is sometimes surprisingly similar – this is called the Nearest Neighbor Rule. The classification problem has now shifted to a search problem: which search algorithm gives the best chances of identifying archive texts that have the same meaning.

In our study, we have used an archive of Dutch pathology reports annotated with SNOMED codes. But the IR approach *itself* is independent of the archive contents. The Nearest Neighbor rule can be used with all sorts of texts in any language, annotated with any coding system. It does not require any modelling of the application domain.

In an earlier report (De Bruijn et al 1997 – chapter 3 of this theis), we described several different search algorithms in detail, and tried to identify which of them gave the best final results. In this paragraph, we confine ourselves to only a short and conceptual overview of the computation of text similarity.

Text similarity for a pair of texts is expressed as a score between *0* and *1*, for *entirely different* up to *most identical* texts. The computation is based on the words that occur in the pair of texts. The more words two texts have in common, the higher this text similarity score.

The contents of a text lies in the words, and in the way the words are arranged. When context is removed, not all of the meaning will disappear due to the intrinsic content of words. This is more so for *content words* (sic) than for *function words*[1]. The importance of a word can be reflected by a weight factor: higher factors for supposedly important words, and lower factors for 'trivial' words such as *the, to*, or *with* are – *and so on*. Words that occur rarely in the entire archive but frequently in the pair of texts, indicate a meaningful connection between the texts. The weight

1) function words are words with predominantly a syntactical function: articles, prepositions, pronouns, conjunctions, and auxiliary verbs. They are like cement for sentence bricklaying.

factors can thus be derived from plain statistics (Sparck Jones 1973). Harter's observation (1975) that function words are fairly randomly distributed in large text collections whereas content words are not, accords with this weighting scheme.

Words can be stemmed into their base forms in order to connect words with the same meaning but with a different appearance. We decided to split word concatenations (very common in Dutch), strip prefixes and suffixes, and make word variations uniform (e.g. 'ae' or 'e'; Latinisms in general). Although there is no hard experimental evidence that supports word unification (but there is none *against* it either) the operation is *intuitively* appealing. A word transformation list may be coupled to the weight factor list – which must be consulted anyway – so transformation hardly burdens the search process.

Instead of using the entire narrative of a report for classification, one may argue to use only the *conclusions*-paragraph. This paragraph, if available, contains all the information that is needed for classification and is usually composed in a 'telegram-style' of writing so that there are fewer function words to spoil the similarity score. On the other hand, the *conclusions* paragraph is often fairly short – one or two lines – so if some function words unfortunately do co-occur, they may cause delusory high similarity scores

The key question now is, whether a computed text similarity score (either for the whole text or for the *conclusions* paragraph only) gives a fair estimation of the real – semantic – similarity between the texts.

3. Experimentation

3.1 Hypotheses

With a collection of reports from two different laboratories, a total of four simulations was run to test the following hypotheses:

simulation 1: test reports from both sites were classified with the pooled archive to test if the search algorithm could correctly classify them.

simulation 2: test reports from site 1 were classified with the archive of site 1 only, and subsequently with archive 2 only. The same was done with test reports from site 2. Thus the influence of site-dependent reporting style was tested: reports from one laboratory may look slightly different than reports from another lab because of different preferences for certain terminology or degree of detail. Since many of the reports were written by two persons – a pathologist and an assistant – the testing of individual differences in reporting style seemed too far-reaching at this stage.

simulation 3: Performance is dependent on archive size, so in order to test this influence an additional simulation was done on two equally large portions of the site-2 archive.

simulation 4: '*Conclusions*-only' retrieval is tested in a separate simulation where only these paragraphs of the reports are used.

Material 3.2

A number of histology reports was collected from two pathology laboratories: 5000 reports from one archive, and 2500 reports from another location. For both locations this was the whole (unfiltered) production of about the first trimester of 1994. A typical report is composed of 80 to 130 words in Dutch medical discourse; this length ranges between 10 and 1300 words. A report contains paragraphs on the source of the material, some clinical data on the patient, the original query of the requesting physician, macroscopical and microscopical observations and the concluding diagnosis. The archive reports were separated from their actually assigned SNOMED codes, so that these codes could form the basis of an independently calculated evaluation metric.

Simulation method 3.3

For the simulations, the 'leave one out' technique was used. This means that one report is lifted from the archive and serves as a trial report. This report is classified by searching nearest neighbors from the remainder of the archive. For the next trial, the former test report participates again in the archive. The method requires no 'learning set'.

Trial reports 3.4

Each of the simulations was repeated 500 times: report numbers 501 to 1000 from each of the two collections served consecutively as trial report. These portions were verified to be representative samples of the collections.

Retrieval technique 3.5

All the archive reports were beforehand transformed into word lists with uniformised words. The words were annotated with weight terms, computed with the following formula:

$$w_{ij} = f_{ij} * \log (D_t / D_i) \quad ;$$

the weight w_{ij} for word i in report j equals the number of occurrences of word i in report j (f_{ij}), multiplied with the log of the ratio between the total number of reports in the collection (D_t) and the number of reports in the collection that contain word i (D_i). This weight term is higher if a word is relatively rare, and drops if a word is more common. Whenever the document collection changes, which happens between some of the simulations in this research, weight factors are recomputed.

The text similarity score for report r1 and report r2, $T(r1, r2)$, is computed with the following formula:

$$T(r1, r2) = \sum_i w_{i,r1} * w_{i,r2} \quad ;$$

If a word occurs in both report r1 and report r2, the weight factors are multiplied and added to the current similarity score. If a word occurs only in r1 then the weight factor for r2 is 0 so that the similarity score is not raised.

Finally, the similarity score is normalised into $T_n(r1, r2)$ which lies between 0 and 1 with:

$$T_n(r1, r2) = T(r1, r2) / \sqrt{(T(r1, r1) * T(r2, r2))}$$

Per trial, the top five most similar archive reports were retrieved by comparing the trial report with each of the reports in the archive collection.

3.6 Method of evaluation

The ideal setting in which a panel of experts would judge the report pairs conflicted with our desire to run a large scale experiment (15.000 report pairs). Therefore a 'silver standard' was devised. The actually assigned SNOMED terms on organ, site and diagnosis were compared per code term, resulting in a calculated SNOMED score. On the basis of the expert ratings from an earlier pilot study (De Bruijn et al. 1997; see chapter 3 of this thesis), the – initially numerical – SNOMED scores were divided into three discrete regions:

- ◊ a higher region in which most of the terms correspond, so that the report is most probably relevant,
- ◊ a middle region for which the result is only partly relevant, dubious or uncertain, and
- ◊ a lower region in which none or only a few of the SNOMED terms correspond between reports: the retrieved report is most probably irrelevant.

For a trial the first nearest neighbor was assessed, as well as the best from the five nearest neighbors. These results were related to the maximum possible score given the archive collection, because for some of the trials the archive contained no relevant solution.

Thus, the measure of evaluation was a 'precision' measure only: 'recall' ('sensitivity' in medical terms) was not under consideration because the hypothesis ('identical descriptions imply the same conclusions') is logically not necessarily true for the inverse ('the same conclusions require identical descriptions').

3.7 Apparatus

For the simulations, we used a standard PC configuration (modest 1994 standards: 486DX33, 8MB). The data required about 9 MB disc space (about 2.5 MB compressed). In this configuration, the simulation per trial took at most 2 minutes.

4. Results

4.1 Simulation 1

The simulation on the pooled archive from both sites yielded the results in table 1. In 627 of the 1000 trials, a relevant classification was given as a first suggestion. In 38 cases, the trial report apparently was a 'rare' case and the archive contained

no relevant other reports at all. The relative retrieval performance was therefore 65.2%. In 87.5% of the classifiable trials, a relevant classification could be found within the first 5 suggestions. In 3.2% of the trials, the first 5 suggestions contained no relevant classification.

The 32 trials in the 'not relevant' category were more closely examined. Sixteen of these cases were exceptional and had no counterpart in the archive. Five cases were partly covered by the suggestions – and should have been placed in the middle region – but the evaluation measure failed to reveal this. Six reports described difficult cases with multiple biopsions and/or an uncommon combination of material and disease – e.g. subcutaneous tissue with lymphocele. The remaining five reports could have been correctly classified given the archive, but were not because correct suggestions were ousted from the top-5 by other achive reports, that were textually more similar. Thus, a skin sample with scar tissue invoked the suggestion 'skin sample with basal cell carcinoma'.

Simulation 2 4.2

Table 2 shows the results when reports were classified with other reports from the same archive (the 'mother'-archive). In 89.5% and 87.8% of the cases respectively, relevant suggestions could be found among the first five suggestions if the archive contained any. The first suggestion was relevant in 65.9% and 65.6% of the classifiable cases. For cross-site retrieval, the number of relevant suggestions dropped and so did the number of classifiable cases. The retrieval performance became 75.3% and 69.6% respectively for the five nearest neighbors.

Simulation 3 4.3

When a smaller archive was used, the number of classifiable cases decreased as can be expected. The number of trials for which relevant classifications were suggested in either the first position or within the first five alternatives, also became less. The relative performance remained about the same – see table 3.

TABLE 1: Simulation 1 - retrieval performance for a pooled archive; 1000 trials from that pooled archive

	number of trials with... (with percentages of max)					
	relevant classification		uncertain classification		no relevant classification	
first nearest neighbor	627	(65.2)	224	(22.4)	149	(14.9)
best of top-5	842	(87.5)	126	(12.6)	32	(3.2)
max	962		38		0	

TABLE 2: Simulation 2 - retrieval performance for searching within the 'mother'-archive and cross-site retrieval

		number of trials with.. (with percentages of max)		
		relevant classification	uncertain classification	no relevant classification
trial reports site 1; archive site 1.	first nearest neighbor	315 (65.9)	122 (24.4)	63 (12.6)
	best of top-5	428 (89.5)	56 (11.2)	16 (3.2)
	max	478	22	0
trial reports site 2; archive site 2.	first nearest neighbor	311 (65.6)	114 (22.8)	75 (15.0)
	best of top-5	416 (87.8)	66 (13.2)	18 (3.6)
	max	474	26	0
trial reports site 1; archive site 2.	first nearest neighbor	202 (44.2)	150 (30.0)	148 (29.6)
	best of top-5	344 (75.3)	110 (22.0)	46 (9.2)
	max	457	43	0
trial reports site 2; archive site 1.	first nearest neighbor	171 (40.0)	148 (29.8)	181 (36.2)
	best of top-5	297 (69.6)	130 (26.2)	73 (14.6)
	max	427	69	4

TABLE 3: Simulation 3 - retrieval performance for a 5000-report archive (site 2) and two 2500-report archives (site 2-a and 2-b)

		number of trials with.. (with percentages of max)		
		relevant classification	uncertain classification	no relevant classification
trial reports site 1; archive site 2.	first nearest neighbor	202 (44.2)	150 (30.0)	148 (29.6)
	best of top-5	344 (75.3)	110 (22.0)	46 (9.2)
	max	457	43	0
trial reports site 1; archive site 2a.	first nearest neighbor	181 (41.6)	164 (32.9)	155 (31.0)
	best of top-5	310 (71.3)	138 (27.7)	52 (10.4)
	max	435	63	2
trial reports site 1; archive site 2b.	first nearest neighbor	189 (43.2)	145 (29.2)	166 (33.2)
	best of top-5	330 (75.3)	126 (25.4)	44 (8.8)
	max	438	59	0

Simulation 4 4.4

If the matching between reports is done on the *conclusions* paragraph only, a relevant classification is given as a first suggestion in 596 trials; in 806 trials a relevant suggestion could be found within the first five alternatives. In 7 trial reports, the *conclusions* paragraph contained no text, and classification is thus impossible. The number of classifiable trials does therefore not add up to 1000. Table 4 summarises the results of this simulation. These results differ significantly with those in table 1 on full-text retrieval ($p<0.001$, paired t-test).

Discussion 5.

In 627 of the 1000 trials with a pooled archive (simulation 1), a relevant solution was given as a first suggestion for classification. Another 215 trials could be classified if the next four alternatives were also considered. So the method could cope with 842 out of 962 (87.5%) classifiable trials, or a total of 84.2%. Sager et al. (1982) reported correct classification in 107 out of 150 cases (71.3%) with their syntax-oriented method. In their research 25 reports did not reach the coding stage; our nearest neighbor method rejected none of the cases. Out of the 125 reports that *were* coded in the Sager study, performance was 85.5% – the generic nearest neighbor method performed comparatively for the first five solutions. The English reports in the Sager study described only head cancer diagnostics, our Dutch histology reports described an unlimited range of examinations. The method itself is independent of the language and there seem to be no reasons to expect poorer results on transfer to English reports or when applying it to other types of narrative text – medical or non-medical.

The number of trials for which the second, third, fourth or fifth suggestion was better than the first one, indicates that the measured textual similarity is not enough to ensure true semantic similarity between cases. The retrieval method did prove very efficient to deliver a low number of highly potential candidates, and a post-processing mechanism might select the best of those candidates. Such a hybrid

TABLE 4: Simulation 4 - 'conclusions' only retrieval on a pooled archive; trial reports from the pooled archive

	number of trials with... (with percentages of max)					
	relevant classification		uncertain classification		no relevant classification	
first nearest neighbor	596	(62.4)	202	(20.3)	195	(19.6)
best of top-5	806	(84.4)	140	(14.1)	47	(4.7)
max	955		38		0	

system could well be the most optimum solution for the classification problem. Finding the 'highly potential candidates' in a large archive is not that easy. For an average report, there are 68 *'relevant'* other reports in the archive, so the 'needle to hay ratio' is 68/7431 = 0.9 %[1].

Classification degrades when a report is classified with the archive of another location (simulation 2). This is the case even after spelling preferences etc. are unified. The results from the simulation seem to discourage the use of a generic archive collection, but definitive conclusions on this matter are postponed until further research is performed. Performances per location were highly comparable, which corroborates the robustness of the method.

Retrieval results did improve slightly with the use of a larger archive (simulation 3). Relative performance remained about the same, and the gain must be attributed to the higher maximum scores that come with a larger archive. However, limitless growth of the archive is not an attractive strategy in striving for better results: a large number of routine cases will clutter the data and make classification slow. It is easy, however, to define the characteristics of a 'rare' case, and stratified growth of the archive could concentrate on these rare cases.

The simulation using the *conclusions* paragraphs only (simulation 4), shows that the results are poorer than with retrieval on the whole text. This means that data storage and retrieval time can be reduced if a summarising paragraph is available for the reports – but at the cost of some reduction in accuracy. The lower results are due to the shortness of the *conclusion* lines – which are often only a few words long: the 'signal' level drops if different words are used to phrase the same conclusion (not all synonyms can be made uniform), and the 'noise' level rises when unimportant words unfortunately co-occur. Note that in practice, the drop from a 87.5% success rate to 84.4% will be experienced as an increase of 25% in debugging the classifications.

The method that is presented here, can be put into use as a module that captures a diagnostic report from the word processor, and suggests suitable classification phrases for the case. This can be done in seconds: the search time that was at most two minutes in these simulations, was decimated after reorganisation of the archive into clusters of reports.

The method itself is independent of the language or contents of the texts: these aspects are determined by the archive collection. Qualities such as accuracy, speed and flexibility can be varied to a great extend by manipulating the archive collection. Results of the experiments were obtained searching an unmanipulated archive; they are expected to be robust if other collections are used. Further exploration of the approach is expected to eliminate the need of coding classification for individual reporters and enhance uniform, objective, consistent and full codes for computerised retrieval in the near future.

---

1) showing a very skewed distribution: the quartile values for the number of 'needles' are 0, 16, 68, 185 and 745.

References

Blois M.S.: Information and medicine: the nature of medical descriptions. Berkeley: University of California Press, 1984.

De Bruijn L.M., Hasman A., Verheijen E., van Nes F.L., and Arends J.W.: Classification of diagnoses that are described in natural language. To appear in: Int J of Technology Management, 1996.

Harter S.P.: A probabilistic approach to automatic keyword indexing, Part 1: On the distribution of specialty words in a technical literature, Part 2: An algorithm for probabilistic indexing. J of the Am Soc for Information Science 1975; 26: 197 - 206 and 280 - 289.

Sager N., Bross I.D.J. Story G., Bastedo P., Marsh E., and Shedd D.: Automatic encoding of clinical narrative. Comput. Biol. Med. 1982; 12 -1: 43-56.

Salton G. and McGill M.J.: Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

Sparck Jones K.: Index term weighting. Information storage and retrieval 1973; 9: 619-633.

Van Rijsbergen C.J.: Information Retrieval (second edition). London: Butterworths, 1979.

4 AUTOMATIC CLASSIFICATION: COMPARISON OF DIFFERENT MODELS

Part 3: Word based model vs. n-gram model

1. Introduction

Narrative text is traditionally given a – strongly redundant – context to allow for easier handling. Library codes are assigned to books so that they will be placed alongside tomes from the same ((sub-) sub-) branch of science – or prose or poetry for that matter. Keywords are added and connections are made through the author's name. An abstract is written to invite people to read the whole text or, for their own sake, deter them from further reading. With the ever growing flood of publications and the electronic availability of them, which are two not entirely unrelated trends, there is greater possibility or even necessity for automatic text processing.

Developments in automatic text processing span several decades to now. To name a few out of the acronym-spangled history of projects: FRUMP (DeJong 1982) applied scripts to interpret wired news bulletins. LSP contained a knowledge structure to classify medical diagnosis reports (Sager 1987). SMART was developed to extract key words from texts for library storage (Salton and McGill 1983). Van Rijsbergen (1979) brought the diverse technical work into a scientific context.

Over the years, a growing interest can be observed for statistical analysis of texts besides a structural one. Probabilistic inferences rather than logical deductions are made by looking not only at the text itself but also at characteristics of other texts – thus including 'observations on the natural environment' of a text. One method from text analysis that uses the environment very directly is the Nearest Neighbor method (cf. Cover and Hart, 1967). This method places a new entry in the same class as the most similar item from a collection. For texts, this would mean that the most similar archive text provides class identifiers or code terms for a new text.

2. Theory

The key in methods like the Nearest Neighbor method is the (morphological) similarity measure or distance measure between two texts, or between one text and a representative of a category of texts. The quality of a morphological similarity measure would lie in the ability of predicting the actual semantical similarity.

If texts are represented in a vector space then a textual similarity measure can be computed as a normalised vector pair inner product (cf. Duda and Hart 1973).

The validity of the computation hinges on the definition of the vector space – i.e. the choice of the axes and the scaling.

Words are the first candidates to span the space; they are after all the molecular content bearers of texts. A disadvantage of *words* is, that morphological varieties such as inflexions, spelling varieties and minor data entry errors may hazard the method's accuracy. A rule based mechanism might correct these varieties but this brings along a complex modelling stage. Character based comparison of words (c.f. Findler and Van Leeuwen, 1979) may restore the connection between morphological varieties but at the cost of burdening the computations.

Damashek (1995) suggested to use character n-grams instead of words for 'gauging text similarity'. N-grams are sequences of n characters, e.g. the 5-grams 'tizen' and 'zen_k' in the string 'Citizen Kane'. N-gram retrieval promises lower vulnerability to data entry errors, spelling varieties, word conjugations, and other morphological varieties. It may even be able to relate texts across languages (the words 'university' and 'universidad' share a number of 5-grams). On the other hand, words that are hardly semantically related may share a treacherously large number of n-grams – e.g.: vaseline/baseline, verse/diverse/ universe.

Chudáçek (1984) defends the use of n-grams for natural language retrieval by describing the statistical properties of tri-grams. The key words from a large collection of texts had almost all at least one identifying tri-gram. The Chudáçek paper unfortunately reports no actual retrieval performance.

3. Experimentation

In order to compare performances of word based classification and the n-gram method, a simulation experiment was run within the context of a larger study (De Bruijn et al. 1996).

Corpus: A subset of 500 texts from a collection of 7500 medical reports of histology diagnostics was used in the simulation. The subset was verified to be a representative sample of the collection, which was in its turn an unfiltered cross-section of histology reporting. A report is typically composed of 80 to 130 words, ranging from 12 to 800 words. Each of the reports was coded in the laboratory with formal classification terms (in the SNOMED classification system). These classification terms did not participate in the retrieval process but were separated from the reports for evaluation purposes.

Simulation method: simulations were run using the *leave one out* method. This means that one report is lifted from the collection and serves as a trial report. From the remaining 499 reports in the collection, the five most similar texts were retrieved. For the next trial, the former test report participates again in the archive. The simulations were repeated 500 times, so each archive report served as a trial report once.

Retrieval method: n-gram based retrieval – the texts were transformed into lists of n-grams, ignoring character-case, numerals and punctuation. A weighting scheme is applied following Damashek's article (1995). This was repeated for 4-grams, 5-grams and 6-grams.

Word based retrieval – the texts were transformed into lists of words, again ignoring character-case, numerals and punctuation. The weighting scheme that was used was adopted from Salton (1983).

Evaluation method: Since the formal codes were known for the reports, the semantical similarity could be estimated independently from the retrieval mechanism by comparing the SNOMED terms in the classifications of test report and retrieved report. Through the results of a pilot experimentation (De Bruijn et al. 1997), the calculated semantical similarity was mapped onto one of three classes:

1. relevant: two reports are alike on material, location and pathology
2. not sure or partly relevant: two reports may be partly alike
3. not relevant: two reports differ too much to be of use in classification.

Results are evaluated in *precision* figures only: *recall* figures are not relevant if the method is used in (Nearest Neighbor) classification tasks.

4. Results

Table 1 lists the retrieval precision of the nearest neighbor with word based and n-gram based similarity calculations. In 273 (54.6%) trials, an archive report was retrieved that had a relevant similarity with the test report. In only 424 trials, a relevant item could be retrieved because in 3 cases, all other collection items were not relevant, and in 73 cases, the semantically most similar report was only partly relevant. Therefore, percentages in tables 1 and 2 are related to these maxima. For word-based retrieval, the relative precision was 64.4%. With n-gram retrieval, the number of trials with relevant nearest-neighbors was significantly lower ($p<0.001$, paired t-test) than word-based retrieval, with 242 (57.0%) for 4-gram retrieval, 246

TABLE 1: relevance of the first nearest neighbor

	relevant (with % of max)		partly relevant	not relevant
maximum	424		73	3
words	273	(64.4)	123	104
4-grams	242	(57.0)	121	137
5-grams	246	(58.0)	124	130
6-grams	249	(58.7)	120	131

(58.0%) for 5-grams and 249 (58.7%) 6-gram retrieval. The differences between 4-gram, 5-gram and 6-gram models were not significant.

In table 2 the precision is displayed for the 'best' solution within the first five neighbors. This is the chance that a user would find a satisfactory solution within the first 5 alternatives. In 369 trials a relevant report was retrieved with the word-based method, which is 87.0% of the maximum. For 4-, 5- and 6-gram retrieval these numbers were 345, 348 and 345 trials respetively, each significantly lower than with word-based retrieval ($p<0.001$, paired t-test). The differences between 4-gram, 5-gram and 6-gram models were marginal and not significant.

Table 3 displays the agreement between methods. The diagonal shows that the word-based method and the 5-gram based method retrieved equally relevant 'nearest neighbors' in 221 + 82 + 83 = 386 trials. In 24 trials, the word-based model retrieved a relevant report while the 5-gram method failed; in 7 trials this was the other way around.

Table 4 shows agreement between methods in 419 trials, in 12 cases the word based model was able to retrieve a relevant suggestion within the first five neighbors while the 5-gram model included only irrelevant suggestions. In none of the trials the reverse was the case.

5. Discussion and Conclusions

In all analyses the word-based method scored significantly better than the n-gram based models. Between the n-gram models (n=4/5/6), no significant differences were observed. In the simulations with a word-based vector space, we measured a relative performance of 64.4% for the first nearest neighbor and 87.0% for the five nearest neighbors. In an earlier large-scale simulation with a 7500-report collection, the relative precision of word based retrieval was essentially the same: 65.2% and 87.5% (De Bruijn et al. 1996). This supports the validity of the present study. There seems no reason to assume that n-based retrieval would increase relative

TABLE 2: best of the five nearest neighbors

	relevant (with % of max)		partly relevant	not relevant
maximum	424		73	3
words	369	(87.0)	91	40
4-grams	345	(81.4)	96	59
5-grams	348	(82.1)	98	54
6-grams	345	(81.4)	100	55

performance with a larger corpus. The agreement between methods indicate that a gain in accuracy can hardly be expected when applying the methods sequentially.

Besides a higher precision, word based retrieval offers a *technical* elegancy: the number of different 5-grams in a text exceeds the number of words with about a factor 6 (own observations). The larger number of n-grams increases storage needs and retrieval time. Only for smaller values of n and a very large and diverse document collection, the number of different words will get higher than the number of different n-grams (Chudáçek).

These experiments indicate that word based retrieval gives better classification results and advantages for implementation in those domains where text errors and spelling varieties are scarce or intercepted, and texts are written in the same language.

TABLE 3: agreement beween word-based retrieval and 5-gram based retrieval - best of the five nearest neighbors

in top-1		5-grams			
		++	+/-	- -	
words	++	221	28	24	273
	+/-	18	82	24	123
	- -	7	14	83	104
		246	124	30	500

TABLE 4: agreement beween word-based retrieval and 5-gram based retrieval - best of the five nearest neighbors

in top-5		5-grams			
		++	+/-	- -	
words	++	333	24	12	369
	+/-	15	60	16	91
	- -	0	14	26	40
		348	98	54	500

References

De Bruijn L.M., Verheijen E., van Nes F.L., and Arends J.W.: Assigning SNOMED codes to natural language pathology reports. In: J. Brender et al. (eds): Medical Informatics in Europe. IOS Press Amsterdam 1996 pp 198-202.

De Bruijn L.M., Hasman A., Verheijen E., van Nes F.L., and Arends J.W.: Classification of diagnoses that are described in natural language. Accepted for publication in Int J of Technology Management, 1997.

Chudáçek J.: Niet-grammaticale verwerking van natuurlijke talen in computers. Informatie 1984 Vol 26-8 pp 594-599

Cover T.M.and Hart P.E.: Nearest neighbor pattern classification. IEEE Transactions on information theory 1967 Vol IT-13 no 1 pp 21-27

Damashek M.: Gauging similarity with n-grams: Language independent categorization of text. Science 1995 Vol 267 pp 843-848

DeJong G.F.: An overview of the FRUMP system. In: Lehnert W.G. and Ringle M.H. (eds), Strategies for Natural Language Processing. Lawrence Erlbaum Associates, Hillsdale, New Jersey. 1982.

Duda R.O. and Hart P.E.: Pattern classification and scene analysis. Wiley, New York 1973.

Findler N.V. and Van Leeuwen J.: A family of similarity measures between two strings. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. PAMI-1 1979 pp 116-118

Van Rijsbergen C.J.: Information Retrieval (second edition). London: Butterworths, 1979.

Sager N., Friedman C., and Lyman M.S.: Medical Language Processing: Computer Management of Narrative Data. Addison Wesley, Reading Mass. 1987.

Salton G. and McGill M.J.: Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

CHAPTER · 5

EXPERT EVALUATION

5 EXPERT EVALUATION

1. Introduction

In the current project, the reporting process for pathology diagnoses is examined. Classification of case reports receives special attention. The classification of a report enables structured computerized storage of the diagnosis and allows powerful search methods to be applied to retrieve data, e.g. for scientific purposes. Classification of the report is now done by the pathologist, but this task has been proven a burdensome one (c.f. Hall and Lamoine 1986) for which the reporter would welcome automatized support. The archiving organisation that receives and manages the data, also advocates automatic classification of diagnosis texts, because as it is, no method is available to check the quality of the classifications or to ensure consistency in classification.

In automatic classification of cases, the point of departure is in fact not the case itself, but its description in a natural language report (with observations, interpretations etc.) or a summary of this. This classification is an abstraction of the report: it will discard some of the details in order to establish a firmer connection with other cases. A good classification in the current context means a complete, correct and concise summary of the findings.

The approach that was chosen here in attacking the classification problem could be regarded as a 'corpus based' method. From a collection of texts, that report is searched that shows the best textual similarity with the text-to-be-classified. If the *best* match is a *right* match, then the classification for the matched report is also adequate for the 'new' case. The method, which will be referred to as 'nearest neighbor' method, may be compared to 'case based reasoning'.

In the implementation of the method, the generic character is protected as much as possible in order to be able to apply the results to other domains than histopathology, other languages than Dutch and other coding systems than SNOMED. No language rules were included: similarity was estimated on the basis of word occurrences only. Weight factors were assigned to words, but they were calculated from frequency counts rather than from discriminating on content. The typical structure of a pathology report, which may have been exploited by including location information in text comparison, was discarded. No direct connection was made between words in the texts and terms in the SNOMED classification system.

By including domain-dependent or language dependent decision rules, the performance of the system may probably be improved significantly. In the

present study, however, we present the results as purely as possible in order to give a realistic image of this system. These results can be considered to give the performance boundaries of the basic implementation of the nearest neighbor method.

In a number of simulations, the implementation already has proven potential (De Bruijn et al. 1996, 1997). For a validation on a larger scale, however, the originally used 'relevance measure' brought along a 'grey zone'. An extensive expert evaluation was desired – and was justified given the previous simulation results. The current paper describes the experiment and discusses the results.

2. Research questions

The central question for this study is: how well do the suggestions for a classification, as determined by the nearest neighbor method, suit the case. This question divides in: how do experts rate the classifications that are suggested for a given report, when do they accept or reject a classification, and could we predict an expert's judgment by comparing the case's original classification with the suggestion. If the latter can be achieved, then ratings can be computed for future comparisons, thus enabling further simulations and estimation of the (notoriously difficult to determine) number of relevant items in a collection – a figure that is needed to come to 'recall' measures.

3. Experimentation

3.1 Material

A collection of reports was acquired from two laboratories: the pathology laboratory of the Academic Hospital in Maastricht supplied 5000 reports, the laboratory from Elkerliek Hospital in Helmond provided us with 2500 reports. In both cases, this was the unfiltered production of histology examinations in about the first trimester of 1994. Appendix 1 of this thesis gives more details on the characteristics of the collection.

From this collection, a total of 240 reports was used in the evaluation experiment. This batch was verified to be a representative sample of the total collection. The 240 reports were divided into six sets of 40 reports: three sets from each of the two laboratories. Each set was to be rated by three referees.

The sets were compiled as follows: starting from a number of random points in the, chronologically sorted, report text file, each report would participate until the sets contained enough reports.

A few reports were excluded:

◊ 2 reports were very long (358 and 599 words, on bone marrow and on mammary tissue) – they would demand a lot of time of the referee.
◊ 1 report was excluded by mistake (nose * choana)
◊ 11 reports were 'too easy': classification was very straightforward with little possibility of variation. These were four reports on naevi, four reports on basal cell carcinomas, two reports on duodenum ('no pathology') and one report on a vas deferens.

The experiment reports contained texts on:

♦ skin tissue (number of reports = 85)
  ◊ naevus (17)
  ◊ verruca (11)
  ◊ basal cell carcinoma (9)
  ◊ multiple samples with different diagnoses (6)
  ◊ dermatitis (5)
  ◊ other (37)
♦ intestinal tissue (57)
  ◊ stomach (26)
    helicobacter pylori (10)
    gastritis / slight abnormality (15)
    adenocarcinoma (1)
  ◊ appendix (4)
♦ intestinal tissue (continued)
  ◊ colon and sigmoid (21)
    inflammation (13)
    other (8)
  ◊ other (6)
♦ gynaecological tissue (19)
  ◊ cervix / corpus (12)
  ◊ uterus (3)
  ◊ other (4)
♦ mammary tissue (11)
♦ bone (9)
♦ vasa deferentia (6)
♦ other (53)

Each of the experiment reports was annotated with five different SNOMED classification phrases: by means of textually comparing the experiment report with the reports in the rest of the site's archive collection (4999 and 2499 remaining reports, respectively), the ten most similar reports were retrieved. From the ten SNOMED classification lines, the first five 'candidate classifications' were selected such that this set contained no duplicates. For 16 experiment reports, the original class-ification line of the report was included in the 'candidate classifications', replacing the fifth suggestion. The order of the 'candidate classifications' was then randomized for presentation. For details on the retrieval method, see (De Bruijn et al. 1997; chapter 3 of this thesis).

Figure 1 gives an example of an experiment report plus 'candidate classifications'.

3.2 Subjects

Seventeen experts participated in the experiment: 14 pathologists and 3 residents from 15 different hospitals. Their average working experience was 16 years (sd 9.9 years), 8 of the pathologists had over 23 years (and up to 29 years) of experience. One of the (very experienced) pathologists volunteered to complete a second experiment session. Subjects were not paid for their participation.

3.3 Procedure

Experts were invited through mail to participate in the experiment; the material for one session was enclosed. If an expert notified us of not being able to participate, or if answers remained forthcoming after reminders, other experts were approached. The response rate was about 65%. Written instructions were included with the experiment material. The instructions mentioned that the results would be used to assess the possibility of automatic classification of reports and it gave a short description of the retrieval method. Subjects completed the rating in their own time by filling out the paper questionnaires. This was estimated to take about one hour of their precious time.

FIGURE 1: example of an experiment report (in Dutch)

AARD MAT.: Abortuscurettement

KLIN.GEGEVENS: Abortus bij amenorrhoeduur van 9 weken.

MACROSCOPIE: Ongeveer 30 cc materiaal. Geen foetale delen aangetroffen. Resten van een vruchtzak als A. Een willekeurige coupe als B. Gedeeltelijk in 2 bakjes.

MICROSCOPIE: Er zijn vliezige delen van een vruchzak aanwezig met rondom chorionvlokken die deels een fibreus stroma tonen en deels een oedemateus stroma zonder vorming van cisternen. De vlokken worden bekleed door een overwegend tweelagige trofoblast
De trofoblast toont een geringe en niet atypische proliferatie. In het vlokkenstroma worden enkele bloedvaatjes aangetroffen, waarin evenwel geen embryonale erythroblasten worden gevonden. Verder necrotisch deciduaweefsel. Geen embryonale delen gevonden.

CONCLUSIE: Abortusmateriaal. Geen embryonaal weefsel aangetroffen.

oordeel:

++	+/-	--	
O	O	O	endometrium * curettement * abortusmateriaal
O	O	O	endometrium * abortus
O	O	O	endometrium * curettage * abortus * zwangerschapsrest
O	O	O	endometrium * curettement * abortus * embryonale delen
O	O	O	endometrium * curettage * abortus

3.4 Rating

Experts were instructed to rate each of the classification lines on a three point scale:

++ the classification is correct, is essentially correct or it can be made correct with a slight intervention;

+/- the classification is partly correct or moderately usable;

-- the classification is incorrect, unusable.

Subjects were asked to keep their interpretation of the scale on a constant level, and to treat the five suggestions per text independently (thus not to rank the suggestions).

3.5 Design

Each of the six sets was rated by three experts in order to compare ratings and to have a 'third vote' if two experts disagreed. In those cases where two experts gave opposite ratings ('++' and '- -'), this was considered so grave that no final rating would be determined, regardless of the rating of the third expert. No final rating was computed either for those cases where one rating was missing and the other two judges disagreed.

4. Results and discussion

For the sake of clarity: in the following paragraphs, the word 'case' will be used to refer to (the rating of) a separate classification line, the term 'experiment report' will be used to refer to experiment reports. Each experiment report therefore produced five cases. When the term 'medical case' is used, the (medical) subject of the report is referred to.

4.1 Consistency and agreement

In 54% of the 1200 cases, all three experts gave the same rating. In 6% of the cases, one rating was missing. In 8% of the cases, it occurred that one expert judged the classification line as '++' while another rated it '- -'. When measured between pairs of experts, average agreement was 0.70 (sd=0.061, range 0.600 - 0.804). 'Disagreement', which is here defined as one expert rating '++' and the other '- -', was 0.04 (sd=0.038, range 0.010 - 0.158).

A Kappa value was computed between each pair of experts: this measures the agreement beyond chance between the evaluations of two raters when both are rating the same object – put on a scale of 0 to 1. For the 18 pairs of experts, an averaged value was found of 0.44 (sd=0,097, range 0,299 - 0,667); values over 0.4 are usually interpreted as a 'fair agreement' (Landis and Koch, 1979).

Reliability was determined for the data: this measures the proportion of the total variation in the data that is explained by variations between objects rather than between observers (cf. Friedman and Wyatt 1997). Reliability equals 1 if all observers agree for all objects. For the observations in this study, reliability was .751; this figure indicates a fair reliability.

Agreement between experts was 'fair', which may be slightly disappointing. Assigning classifications to medical cases is usually thought of as being a straightforward task in which the rigid classification system creates a black and white situation. These results indicate that there is a *shaded area* caused by (subjective) interpretation of the SNOMED classes. The task in this experiment, however, is a somewhat artificial situation: the classification of an other – more or less differing – medical case must be judged, and a rating must be given to the level of discrepancy with the true or optimal classification. Part of the deviation in agreement may therefore lie in the subjective use of the rating scale – this will be expressed mostly as one expert judging '+/-' and the other expert judging '+ +' or '- -'. However, a stronger indication that experts are not always as consistent as was expected in rating tasks such as these is expressed by the 96 cases that were given a '+ +' rating by one expert and '- -' rating by another.

In conclusion: rating classification lines is not a straightforward task, but it rather causes variance in the judgments. This variance also indicates that subjective latitude can be expected when experts must assign appropriate classification terms to a report.

Expert subjectivity: rating of original classification 4.2

For 16 experiment reports, the original classification line of the report was inserted among the five suggestions. For 47 experiment reports, one of the suggestions was identical to the original classification line. These 63 cases were further analysed. A unanimous '+ +' expert rating was found in 29 cases. In 15 cases, two of the experts rated the classification as 'partly correct or moderately usable' ('+/-'). In 13 cases, one or more of the experts rated the original classification as '- -'. Table 1 lists these observations with the experts' judgements.

For the experiment reports of his set, one pathologist (on his own initiative) wrote out two full classification lines: one classification line (C_1) as he would assign it to the set himself, and a second classification line (C_2) in the manner how it is formally required, but 'which is not always followed in practice'. If C_1 and C_2 are compared, they agree on terms[1] by .84 (sd=.13, range=.55 - 1.0), if they

1) the similarity between classification lines is computed through the 6-character SNOMED codes. The codes in the two code lines are compared one by one; two strings being the same increase similarity with 1, two strings being completely different increase similarity with 0, and if two strings differ on a deeper level in the hierarchy then similarity increases with a fraction of 1. After comparison, the similarity is normalized. See also section 5.3 of this chapter.

are compared with the original classification (C_0) as it was given at the time of diagnosis, C_1 agrees with C_0 by .75 (sd=.18, range=.35 - 1.0), and C_2 agrees with C_0 by .81 (sd=.16, range=.35 - 1.0). In less than 25% of the cases, C_0 was identical to either C_1 or C_2. These results were obtained from outside the controlled design of the experiment and should therefore be considered critically.

In another experiment that was run within the same project (Wieggers et al. 1997), experts were asked to *compose* classification lines for report texts rather than judging lines that are presented. Comparison between the experts showed that their classification lines corresponded with .77 (sd=.15, n= 120 (4 experts x 30 reports)). Comparison of each expert with the original classification showed a correspondence of .78 (sd=.16, n= 120). In less than 10% of the cases, a pair of pathologists wrote exactly identical classification lines.

These figures show that for a single medical case, SNOMED classifications can be expected to differ to a rather large extent when written by different pathologists.

As was discussed in section 1, the main task of this experiment describes a somewhat artificial situation: judging the classification of a differing case. The ratings of the 63 original classifications however, involve a different, more primary task: judging classifications that were given to that case *in reality* – in a clinical situation so that the classification should be in principle correct. In about half of the cases, the outcome indicates that either (1) the expert who assigned the classification line to the report in the first place, must have had an 'odd' view of classification, or (2) the expert(s) who rated the classification line did.

TABLE 1: expert ratings for the original classification lines

expert ratings			number of observations
++	++	++	28
++	++	missing	1
++	++	+/-	11
++	+/-	+/-	10
++	+/-	--	3
+/-	+/-	--	5
+/-	--	--	2
++	++	--	2
--	--	--	1
total			63

With respect to the deviations in judgments, two possibilities may apply:

1. for each medical case, there is one and only one suitable SNOMED classification, but in practice, incomplete, overcomplete or incorrect classifications are sometimes given to reports. A judge in experiments will detect those errors.

In an evaluation task – such as this – one makes slips less easily than in a generation task (such as classifying cases). A case is classified by *one* expert, and the classification is judged by *three* experts in this experiment. The classifying expert worked in a daily setting, whereas the judgement took place within an experiment setting and could therefore show an additional advantage by Hawthorne effect[2].

This underlines the importance of automatic classification (-support): errors are being made and as it is, there is no instrument to identify them. Automatic classification could verify the correctness of classification lines, or participate in the classification process itself. Note that a corpus-based method such as presented here, uses past classifications which are therefore not entirely reliable. These methods would need an offset time after which incorrect classifications have expired from the archive.

2. for each medical case, more than one SNOMED classification may be suitable – one always has to work with subjectivity in assigning classifications. A classification may be correct in one expert's eye but needs improvement in the eye of another expert.

If this is the case, differences in interpretation of the rating scale will have a greater influence. One judge may indicate that she accepts the given classification even though she would have classified the case differently herself, another judge may give a negative rating if the classification does not completely concord to his own point of view. The variations in classification lines when they are written by different pathologists, support this hypothesis. For a single case, several different classification lines may be given, all correct but some more correct than others.

This, again, underlines the importance of automatic classification, even if it 'only' *supports* the classification: in a method like this, the reporter is confronted with classification lines that were written by other reporters. Alternatives are presented, and (hidden) communication like this between reporters should reduce the individual differences in classification.

The hypotheses cannot be extensively tested with the relatively small number of observations, and the truth probably can be found somewhere in the middle.

---

1) A person's performance in experiment situations tends to be better than in daily routine. The name of this phenomenon was derived from the Hawthorne plant of Western Electric where production increased when workers became aware of participating in an experiment.

FIGURE 2A: text similarity vs. expert rating. Vertical variation within classes is added for the sake of clarity

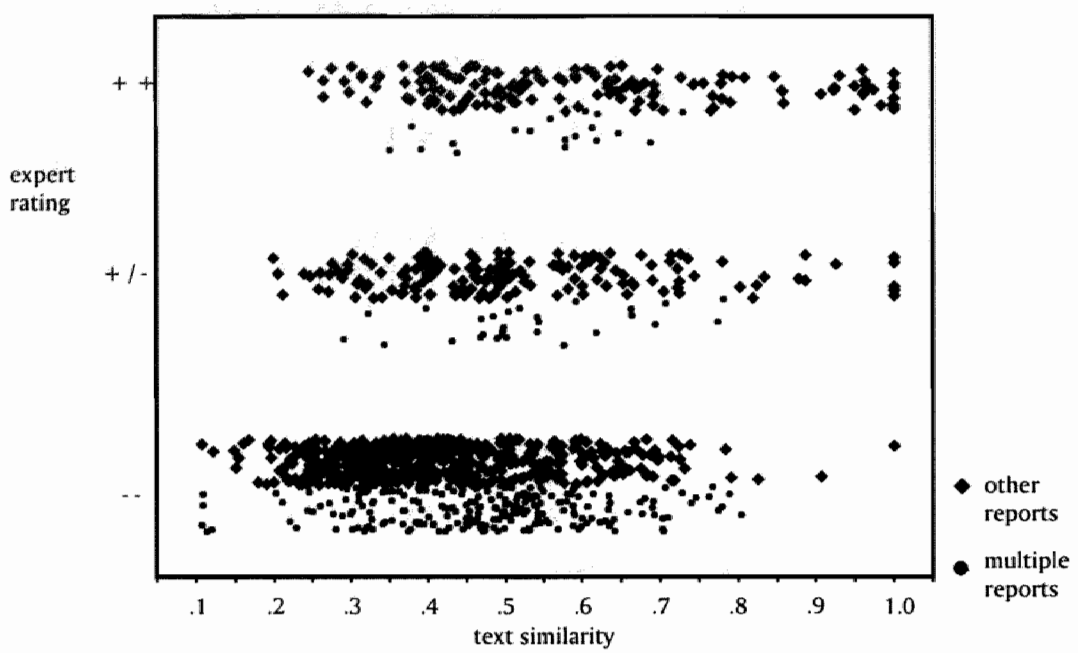
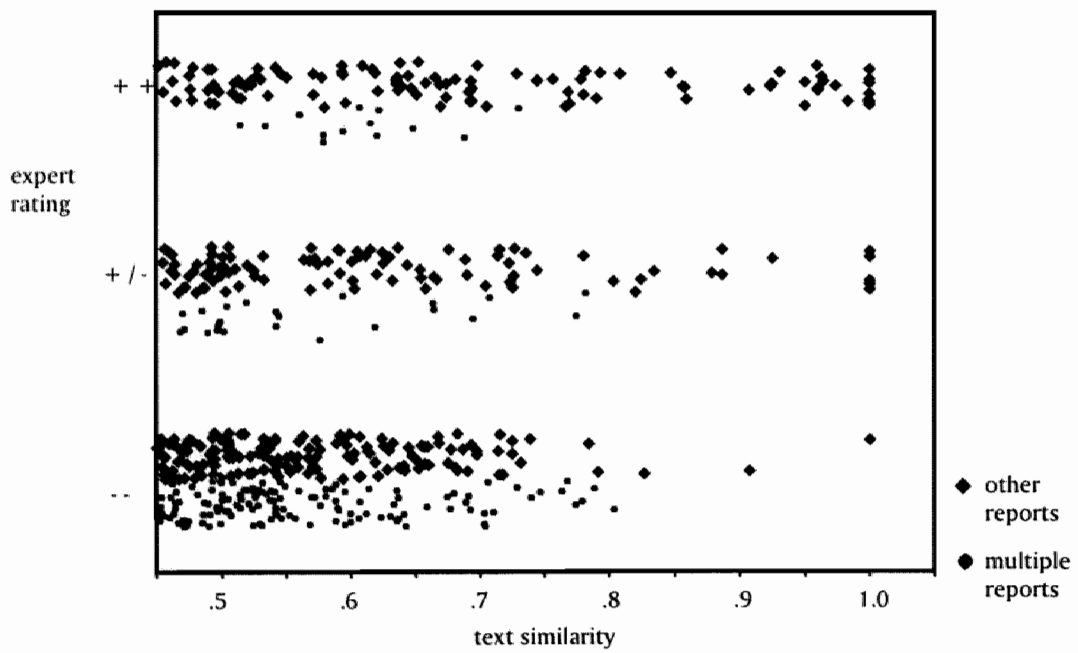


FIGURE 2B: text similarity vs. expert rating for values > .5. Vertical variation within classes is added for the sake of clarity



Above observations – especially since they were gathered from data from the clinic – show clearly that in the current situation the level of objectivity and standardisation for which classification systems are actually designed, is not attained. Automatic classification (-support) may be a valuable tool in reaching the desired level of uniformity.

A look inside the reports 4.3

The classification method assumes that the (textually) most similar report from a collection gives the best prediction for classifying the case. If the textual similarity increases, the probability of the *best* prediction being a *right* prediction, increases. The runners-up can also give a good classification if they are close behind the nearest neighbor, and/or if they too have a high textual similarity with the report-to-be-classified.

It may happen that two reports describe a similar case, but the textual similarity comes out low because different words are used. Such a case will probably not be retrieved by the method. If the method were to be used to extract *all* relevant cases from a collection (perfect 'recall') and *nothing but* all relevant cases (perfect 'precision') then this effect would count as a disadvantage. For classification however, one retrieved correct report suffices. The logical asymmetry is not as paradoxical as it may seem: a high textual similarity implies identical classification – identical classification does not imply a high textual similarity.

Figures 2a and 2b show the expert ratings against the textual similarity measure. Figure 2b displays the textual similarity measure > .50 in order to display the higher values better. Multiple reports are marked distinctly from other reports for reasons that are explained below. The markers on the extreme right represent those cases in which the original classification was rated. The textual similarity with the source report naturally is 1 on a scale of 0 to 1.

The graph shows that with a higher text similarity the chances are higher that the expert rating is '++'. Still there is a number of cases for which a high text similarity harvested a low expert rating. These cases received special attention in further analysis: the individual cases were scrutinized, and described in the following paragraphs.

Model performance 4.3.1

For 52 out of 240 experiment reports the classification mechanism performed well. The first suggestion and close runner-ups were rated as '++' by the experts, lower rates were only given if the textual similarity had dropped considerably.

4.3.2 Multiple reports

These gave problems: a multiple report describes an examination of sections from two or more different locations; typical combinations are stomach + colon, stomach + oesophagus, appendix + gall bladder, and endometrium + endocervix.

For 42 experiment reports, the report described a multiple section. In 16 instances, one of the classifications was rated '++', in 9 instances the best classification was rated '+/-' and in 17 instances all classifications were judged '--'.

In 87 cases (73 different experiment reports), the experiment report was singular but multiple classifications were retrieved. In 77 cases, the rating '--' was given, in 7 cases '+/-'. In only 1 instance, the case was rated '++' and in 2 cases a final expert rating could not be given.

A multiple report may resemble another multiple archive report and therefore retrieval may be successful. However, the retrieval method discards context information, so that chances are that the final diagnoses are swapped between the different topographies. For instance, a report that describes

> skin sample from the chin with a basal cell carcinoma *and* a skin sample from the belly with a naevus,

may be highly similar to another report that describes

> skin sample from the chin with a naevus *and* a skin sample from the belly with a basal cell carcinoma.

If the resemblance with a retrieved archive report is lower, part of the classification may still be useful, but experts gave low ratings to these cases. A similar argumentation goes for 'singular' experimentation reports that raised a multiple report from the archive – in some of these cases, part of the classification was correct but the suggestion was discarded by the expert on the basis of the surplus part.

In order to cope with these cases, the retrieval method should be enhanced with a routine that separates paragraphs or sentences on the different slides (a concession towards practicality but away from genericity) or by using a retrieval method that keeps part of the context intact – e.g. by including word pairs (words that are adjacent, or in direct proximity of each other) in the computation – which would make the method more complex.

4.3.3 Rare cases

For 42 experiment reports, additional manual searching revealed that none of the classifications in the collection suited the experiment report. For three of these reports, an extremely low text similarity figure was found for the first nearest neighbor (.120, .217, .223 – where the average text similarity for the first retrieved archive report was .53 (sd=.16)). For the other reports, the first nearest neighbor described a case with a similar organ biopsy but with another diagnosis or other diagnoses.

A larger collection is likely to reduce the chances of a case not having a counterpart in the archive.

Skin reports 4.3.4

For 25 experiment reports, the experiment report describing skin tissue was classified with correct diagnoses but giving a different location of the skin, and for that reason given a '- -' rating by the expert for all suggestions.

The classification method relies most on the observations in the 'microscopy' paragraph in the report. On a microscopical level, skin is skin, so archive reports are rather retrieved on the basis of pathological status than on the exact location. A procedure rule could intercept the 'location' problem: if the report under consideration is a skin biopsy, then the exact location can in most cases be found in the 'aard materiaal' line. The classification line can then be adapted accordingly. Again, this is a concession towards practicality but away from genericity.

Table 2 displays the distribution of the 1200 cases in four groups: singular and multiple skin biopsies, singular and multiple non-skin biopsies. The outcome shows that multiple cases (those cases where the report or the suggestion is multiple) are rarely rated '++' and the retrieval method would only work for singular reports, with adaptation required when processing skin biopsies.

In one case, (huid * eruptie), the original classification of the report was rated '- -' by all three observers, possibly because neither the location nor the excision method was indicated in the classification.

Discriminant terms 4.3.5

In most situations, the meaning does not depend on a single word because additional observations (expressed in other words) substantiate the final conclusion, and aid the Nearest Neighbor method. In a few cases, the conclusion *does* depend on that single word, and leaves the method powerless. These cases are fairly easily defined and may be solved by an additional (NLP-

TABLE 2: Distribution of cases in four classes and outcome (total collection)

	other					skin biopsy				
	++	+/-	- -	?	tot.	++	+/-	- -	?	tot.
singular	107	93	285	61	546	46	67	218	26	357
multiple	13	24	175	17	229	5	4	51	8	68
										1200

like) analysis of the phrase around such a key word.

In 14 cases (7 experiment reports), a classification was judged '- -' because a single discriminant term in the otherwise textually very similar archive report caused a crucial difference in the classification line: sidedness left vs. right (7 cases), the degree of dysplasia or bacterial colonisation (4 cases), resection edge free vs. not free (2 cases) and (no) ferritine depletion (1 case).

Such problems are highly specific but may easily be solved by a small number of interception procedures. A text can be scanned for words indicating left/right/double sidedness and the classification line can be altered accordingly. Parsing rules or interception procedures may only be needed in a limited and well-defined set of occasions, and easy to implement, but they are again a violation of the genericity ideal.

4.3.6 Treacherous textual similarity

On two occasions, a medical case with a certain disease retrieved a nearest neighbor which was suspected for that disease but which showed not enough evidence to diagnose it as such. These were a skin report with a suspicion for compound naevus but finally diagnosed as a dermal naevus, and tissue from the ear with (suspected) chondrodermatitis. The method naturally retrieves such a report. Since these are fairly rare occasions, they may be intercepted when, apart from the first nearest neighbor, also other near neighbors are considered.

Concluding: the cases in which a high textual similarity did not guarantee a positive expert rating are explainable. Solutions are possible that intercept errors, e.g. additional scanning on left/right sidedness or the topography of skin biopsies. These solutions are further discussed and evaluated in chapter 6 of this thesis. The results of this experiment, however, give an idea of the 'Achilles heels' of the Nearest Neighbor classification rule.

4.4 Judgement of the experiment sets

A final judgement could be determined for 1088 of the 1200 cases. In the other 112 cases, judgments varied between '++' and '- -' or differed while the third rating was missing. In 46 cases, two experts agreed on '++' (23) or '- -' (23), but the third expert gave an opposite rating so that these cases were not included. The ratings were distributed as follows:

++	:	171
+/-	:	188
- -	:	729

Per experiment report, five classification lines were suggested and rated. If the classification system is tight, only one of the five different classifications will be correct, and this distribution is to be expected.

For 107 of 240 reports, at least one of the suggestions was rated '++' (excluding the inserted 'original suggestions'); for 57 reports the best suggestion needed adaptation (rating '+/-') and for 75 cases, none of the classification lines was rated even partly adequate ('- -'). For one report, no final score could be determined for any of the suggestions.

These results compare to earlier simulations (De Bruijn et al. 1996) as follows: the expert ratings in the experiments reported in this chapter (which should be considered as a 'golden standard') indicated a success rate of 44.6% for the top-5 suggestions whereas the simulation experiments came out on 84.2%. In these simulations, results were evaluated through a similarity figure which was computed between the original classification and the suggested classification. The difference between the experiments puts question marks at the Silver Standard definition. Several notes should be made, however:

note 1: the designs of the experiments differ on a number of points. The Silver Standard that was used, was based on a pilot experiment in which an expert rated the correspondence of pairs of medical cases by reading the full text reports. In the present experiment, the full text of the archive report was not available for the referee – only the classification line. In the simulation experiments, the first five suggestions were evaluated; in this experiment, duplicate suggestions were removed so that the rated suggestions originated from the top-10.

note 2: eleven easy reports were excluded from the set of experiment reports, as well as three difficult reports that may have been rated as '- -' (see section 3.1). The corrected outcome would be: (107 + 11) / (240 + 14) = 46.5%.

note 3: As was indicated in section 4.3.2, multiple reports proved to give problems in this experiment. Out of 297 cases where the experiment report or the retrieved classification was multiple, 226 were rated '- -' and 18 were rated '++'.

The outcome for the singular reports rated as '++' was 90 / 195 = 46.2%; the 'Silver Standard' judgment for the same cases would be 81.5%. Combined with the correction for eliminated reports (*note 2*) the outcome would be 51.5%.

note 4: reports seemed to have been severely judged.
First, there is the matter of 'contrast'. In the absence of 'really bad' suggestions, suggestions that are not totally correct but that do have *some* relevance may be judged as '- -'. In virtually all cases, the classification had *something* to do with the report – a concept akin to 'relevance' that is sometimes named 'aboutness'. In only 2.9% of the cases, the SNOMED terms of the original classification differed

more than 80%[1] from the suggested classification; in only 6.6% of the cases the difference was >70%; in 76.3% of the cases, the classification lines agreed by more than 50%. In an experiment situation there is no use in offering the expert classification lines that have *nothing* to do with the case because even a layman knows the outcome of these trials beforehand, and you needlessly use the expert's time (= experiment resources). The experimenter even risks to lose the subject's interest or perhaps all cooperation. These are inherent disadvantages attached to the evaluation method.

As was discussed in section 4.3, skin biopsy reports were strictly judged because of a discrepancy in the exact source (location) of the material. A classification differing in 'location' only still got a decisive '- -' judgment of the experts in 24 (out of 74) (singular) skin sections. In the earlier simulation experiment these cases would be placed in the '++' category: 'a relevant classification – to be corrected with a minor action'.

Similar arguments can be put forward for a number of 'dichotomous' observations: sidedness 'left' or 'right' (or 'bilateral' or 'unknown'), resection edge (not) free – and also the degree (1, 2 or 3) for tumors or bacterial colonisation. This was crucial for 7 experiment reports.

It is not my point to trivialize the importance of correctly coding the topography, and the experts have pinpointed a weak spot of the Nearest Neighbor method with their ratings. A solution lies in adding Natural Language Analysis procedures. The exact location in a skin report is most often given in the '*Aard Materiaal*' paragraph of the pathology report, and repeated in the first phrase of the '*Conclusie*' paragraph. Only a few lines of program code added to a descendant of this classification program scanned the text for the words 'rechts', 'rechter', 'links', 'linker', 'dubbelzijdig', and 'beiderzijds' and then adapted the suggested classification line for sidedness. Other enhancements were also made, which will be discussed in chapter 6 of this thesis. In this evaluation, however, we wanted to use the Nearest Neighbor method as purely as possible without flattering the results with (albeit useful) ad-hoc solutions.

Upon successful implementation of the enhancements, the corrected outcome is estimated to be 66.5%.

note 5: A more severe judgment does not need to be a negative thing. In earlier chapters I argumented that the basis of the method is the archive collection – if the collection does not contain a second identical medical case, the method can search but will never be successful. That is the reason why in previous experiments results were related to the maximum obtainable score. It appeared then that in 96.2% of the cases a sufficiently similar case was available in the collection – where *sufficiently similar* was defined through the 'Silver Standard'.

1) See section 5.3 of this chapter.

If the golden standard is more severe, this 'retrieval ceiling' is likely to get lower. In evaluation with expert ratings, there is no practical means to determine the 'ceiling' (we did not even *try* to persuade experts in rating 7499 classification lines per experiment report). This is a recognised constraint of evaluation with expert ratings.

Without such a ceiling, correcting the outcome percentage is not possible. However, in section 5 an alternative method will be shown to determine it.

Even though a number of factors have put the results of the expert ratings in a perspective, these ratings show a lower success for classification than was estimated in earlier simulations (De Bruijn et al. 1996). The main factor seems to be that judgments were severe, rating suggestions as '- -' when they differed from the original classification in only one or two (apparently crucial) classification terms. For a number of reports, the crucial difference can be bridged by additional checking of the report or the suggestion for the absence or presence of a limited number of diagnostic 'parameters'. Separate processing for the different segments of multiple reports may give more realistic suggestions for classifying the multiple cases. For some reports, the archive did not contain classification lines that satisfied the (severe) judges' requirements – and if there is nothing to be found, the searcher cannot be blamed. Extension of the archive collection may counterbalance this constraint.

5. Towards a Silver Standard

A disadvantage of using expert ratings as a *golden standard* is that items that are *not* rated cannot be included in the analyses. With a corpus based method such as the Nearest Neighbor method, the retrieval mechanism should be evaluated by relating the (retrieved) outcome to the remainder of the corpus. Effectiveness cannot be estimated without knowing details on the contents of the corpus. A *Silver Standard* that is determined through computations on the classification line may be a valuable estimate in rating those cases that the expert has not seen.

The original Silver Standard as was presented in the pilot study (De Bruijn et al. 1997), consisted of a string comparison figure between the coded classification line of the test report and that of the retrieved report. The classification lines do not participate in the retrieval process, and can be used for independent evaluation. The comparison of classification lines obviously can only be used for evaluating simulations, because in a practical setting, the classification line of the report-to-be-classified is not yet available.

In the following paragraphs, a new Silver Standard is obtained using the collected expert ratings.

5.1 The original Silver Standard

The original Silver Standard from the pilot study (De Bruijn et al. 1997), consisted of a simple string comparison between the codes in the classification lines of the test report and the retrieved report. Any 6-character SNOMED-code that appears in both classification lines, raises the similarity measure. The similarity measure $S_{1,2}$ is computed with the formula

$$S_{1,2} = \sum f_{1i} f_{2i} \; / \; (\sum f_{1i}^{\,2} \; \sum f_{2i}^{\,2})^{\frac{1}{2}} \qquad (1);$$

if term i occurs once in string 1 ($f_{1i} = 1$) and once in string 2 ($f_{2i} = 1$), then $f_{1i} f_{2i} = 1$ and the similarity measure is raised. The denominator of the formula normalizes the similarity measure. Identical classification lines give $S_{1,2} = 1$, for entirely different classification lines, $S_{1,2} = 0$ (see also Findler and Van Leeuwen 1979).

On the basis of the results of the pilot experiment, the similarity measure was divided into three regions:

++	similarity measure > .56	reports are likely to describe similar observations;
+/-	similarity measure $\in$ [.32, .56]	reports are likely to describe partly similar observations, or indicates uncertainty in similarity;
--	similarity measure < .32	reports are likely to describe dissimilar observations.

Table 3 shows that the original Silver Standard agrees poorly with the expert's rating from the large scale evaluation experiment. The table illustrates that the experts rated the suggestions much more severe, or in other words, the Silver Standard allowed more deviation from the original classification in rating it as 'relevant'. There were great differences in design between the pilot experiment and the large-scale expert rating experiment, so a differing outcome may be

TABLE 3: correspondence between original silver standard and expert rating

		expert rating			
		++	+/-	--	
original silver standard	++	156	130	295	581
	+/-	13	46	241	300
	--	2	12	193	207
		171	188	729	1088

		sd
Kappa	.140	.014
Spearman Correlation	.395	.023
agreement: (395 / 1088)	.36	
disagreement:	.273	

expected. These observations lead to the conclusion that the Silver Standard needs adaptation in order to give more dependable results in further simulations.

Original Silver Standard, optimized 5.2

The original Silver Standard can easily be tuned, namely by shifting the thresholds that define the difference between the categories '++', '+/-' and '- -'. With a shift of thresholds, the column totals of the cross table remain the same, but the distribution over the rows varies. Threshold values were determined such that row totals corresponded optimally with column totals, which gave the following values:

++ similarity measure > .83
+/- similarity measure ∈ [.67, .83]
- - similarity measure < .67

Table 4 shows that these boundaries give a better correspondence with the expert's rating.

As was argued in section 4.3.2 of this chapter, two multiple reports may have many code terms in common but still differ semantically if code terms apply to different sections of the report. Multiple reports should therefore be handled with care. The same argument applies for 'skin' reports, which are sometimes severely judged (see section 4.3.4 of this chapter). Finally, a distinction is made for reports from different laboratories. Tables 5a and 5b show that optimization of the Silver Standard gives different thresholds for the two laboratories for singular non-'skin' reports. The expert ratings for test reports in the 'Maastricht' corpus seem to be less well predictable than for the 'Helmond' collection. The reasons for this are unclear.

The optimized original Silver Standard still gives outcomes that differ much from experiment ratings. A better definition of a Silver Standard seems achievable, and will be explored in the next paragraphs.

TABLE 4: correspondence between original silver standard (optimized) and expert rating

		expert rating			
		++	+/-	- -	
original silver standard	++	92	43	37	172
	+/-	40	47	106	193
	- -	39	98	586	723
		171	188	729	1088

		sd
Kappa	.332	.025
Spearman Correlation	.485	.028
agreement:	.67	
disagreement:	.070	

5.3 Alternative Silver Standards: Silver Standard II

SNOMED is a hierarchical and multi-dimensional coding system. This structure can be used for refining the similarity measure between code lines. SNOMED classifications in the collections use 6 dimensions or axes (after Gantner 1979): Topography, Morphology, Procedure, Etiology, Function, and Disease. Each of these axes has a hierarchical character which is visible in the formal code that is associated to the concept – a six-character code in which the first character describes the axis. The word 'head' (code TY0100) is the *parent* of the term 'forehead' (TY0110), *sibling* of 'neck' (TY0600) and *child* of the more general concept 'head and neck' (TY0000). A parent/child relationship on a certain level in the hierarchy is represented by the parent having the digit '0' on the corresponding place in the code string, and the child having another character there ($\in$ [1..9], in general). A single concept may be connected to a concatenation of codes: 'appendicitis' has a topographical as well as a morphological component (T66000M40000). See also section 3.4 of this thesis' introduction.

TABLE 5A: optimized original silver standard and expert rating (Maastricht - excluding multiple and skin reports)

		expert rating			
		++	+/-	--	
original silver standard	++	25	13	11	49
	+/-	10	10	29	49
	--	10	12	122	144
		45	35	162	242

boundaries:	0 - .66 - .78 - 1		
Kappa		.343	(.050)
Spearman Correlation		.515	(.057)
agreement		.65	
disagreement		.087	

TABLE 5B: optimized original silver standard and expert rating (Helmond - excluding multiple and skin reports)

		expert rating			
		++	+/-	--	
original silver standard	++	45	9	6	60
	+/-	11	23	27	61
	--	6	26	90	122
		62	58	123	243

boundaries:	0 - .55 - .76 - 1		
Kappa		.438	(.048)
Spearman Correlation		.609	(.048)
agreement		.65	
disagreement		.049	

The hierarchical character of the classification system can be incorporated in the string similarity computation by the use of a fraction of 1 that matches the level in the hierarchy on which codes correspond:

◇ codes that correspond up to the 6th (deepest) level of the hierarchy	$f_{1i}f_{2i} = 1$
codes differing on the 6th level with a parent/child relationship	$f_{1i}f_{2i} = .9$
◇ codes that correspond up to the 5th level of the hierarchy	$f_{1i}f_{2i} = .8$
codes differing on the 5th level with a parent/child relationship (e.g. T77100 - glandus prostaticus / T77120 - ductus prostaticus)	$f_{1i}f_{2i} = .7$
◇ codes that correspond up to the 4th level of the hierarchy (e.g. T77120 - ductus prostaticus / T77140 - musculus prostaticus)	$f_{1i}f_{2i} = .6$
codes differing on the 4th level with a parent/child relationship	$f_{1i}f_{2i} = .5$
◇ codes that correspond up to the 3rd level of the hierarchy	$f_{1i}f_{2i} = .4$
codes differing on the 3rd level with a parent/child relationship	$f_{1i}f_{2i} = .3$
◇ codes that correspond up to the 2nd level of the hierarchy (the 2nd level has no parent/child relationship)	$f_{1i}f_{2i} = .2$
◇ codes that correspond up to the 1st level of the hierarchy (this level represents only the axis to which the term belongs)	$f_{1i}f_{2i} = 0$

A similar computation as formula (1) yields a single, normalized similarity figure. Tables 6 and 7 show the relation between this similarity figure and expert ratings – given for the entire test collection (table 6) and the singular, non-'skin' reports from Maastricht and Helmond (tables 7a and 7b). These tables show slightly improved results with tables 4 and 5, but these differences are not significant. This similarity measure may be applied, but does not give more precise results than the optimized original Silver Standard.

Alternative Silver Standards: Silver Standard III 5.4

The expert's judgment may be better modelled by assessing each of the dimensions (or axes) of the SNOMED classification system separately. Therefore, separate similarity figures were calculated for the T-axis, P-axis, M-axis, D-axis, E-axis and F-axis in the same manner as in the previous paragraphs. The parameters 'skin' and 'multiple' were added. These parameters were used in a

TABLE 6: correspondence between optimized silver standard II and expert rating

		expert rating			
		++	+/-	--	
original silver standard	++	98	46	40	184
	+/-	38	41	96	175
	--	35	101	593	729
		171	188	729	1088

Boundaries:	0 - .75	- .865 - 1
Kappa	.341	.025
Spearman Correlation	.502	.028
agreement:	.67	
disagreement:	.069	

(non-parametric) discriminant analysis. In the results, the parameters representing the axes D, E and F proved not significant. The plot of the two canonical discriminant functions indicated that only the first canonical function discriminated between groups, and the second function gave no further discrimination. The single equation that was returned by the analyses, was:

Rating Estimate $= .6 * Sim_M + .3 * Sim_T - .17 * P_{Mult} - .14 * P_{Skin}$ (2);

thus the estimate for the expert rating varies positively with the similarity between codes on the T-axis (Sim_T) and the M-axis (Sim_M), and depends on whether the case is 'multiple' or a 'skin'-report (P_{Mult} , $P_{Skin} \in (0, 1)$). The parameters showed to be robust when estimated with random subgroups, and also when estimated with the singular non-skin reports.

This discrimination variable was mapped to three regions (corresponding with '++', '+/-' and '- -') with the following thresholds:

$< .53$:= '- -'
.53-.75:= '+/-'
$> .75$:= '++'

TABLE 7A: optimized silver standard II and expert rating (Maastricht - excluding multiple and skin reports)

		expert rating			
		++	+/-	- -	
original silver standard	++	22	14	11	47
	+/-	14	6	15	35
	- -	9	15	136	160
		45	35	162	242

boundaries:	0 - .72 - .865 - 1	
Kappa	.356	(.049)
Spearman Correlation	.563	(.055)
agreement	.68	
disagreement	.083	

TABLE 7B: optimized silver standard II and expert rating (Helmond - excluding multiple and skin reports)

		expert rating			
		++	+/-	- -	
original silver standard	++	46	9	6	61
	+/-	13	21	26	60
	- -	3	28	91	122
		62	58	123	243

boundaries:	0 - .63 - .83 - 1	
Kappa	.438	(.048)
Spearman Correlation	.637	(.045)
agreement	.65	
disagreement	.037	

Table 8 displays the cross table between the experts' ratings and Silver Standard III. The table shows improved accordance with experts compared to the original Silver Standard and to Silver Standard II; the Kappa was significantly higher ($p<.001$). Tables 9a and 9b show these figures when used on the singular, non-skin reports from the two collections.

The 15 cases in the upper-right corner included three cases where the *original* classification was rated '- -' by the experts and twelve cases where the classification lines differed in not more than one classification term. In three of those cases, the differing term involved left/right sidedness, in three other cases, the difference involved a term on the E-axis.

The 23 cases in the lower-left corner of table 8 were given a '++' rating by the experts despite differences in classification lines between the original and suggested classification. They included eight cases where differences in coding for the M-terms *and* the T-terms were allowed by the experts – two of those involved 'skin'-reports. In seven cases, classification lines differed on M-terms alone – two of those involved 'skin'-reports. For one case (a 'skin'-report), the difference was due to the T-terms alone. Seven cases were multiple reports that were rated '++' by the expert, three of these were 'skin'-reports.

If Silver Standard III is compared with each of the individual experts, an average Kappa value is found of .344 (sd=.092 - range .228 - .479), average agreement is .644 (sd=.035 - range .575 - .705) and disagreement averages to .055 (sd=.034 - range .005 - .126). These figures indicated that this Silver Standard model as yet offers the best possibility of estimating the correspondence in meaning between report pairs. Values even approach those as measured between experts.

Silver Standards – Conclusion 5.5

Various ways of coming to a Silver Standard were explored. A disciminant function was determined, which proved better than the original Silver Standard that was developed from pilot experiment data. With a good Silver Standard, additional cases may be rated so that large scale simulation experiments can be run, and with a higher level of reliability than the original Silver Standard allowed.

The Silver Standard models do not always agree with the expert's ratings. It should be noted that experts did not agree among themselves in rating classification lines, so their behaviour has a certain level of variance that cannot be modelled. Furthermore, the Silver Standards are computed on the basis of the original classification that, as was shown in section 4.2, also include a degree of uncertainty.

For evaluation of large scale simulations, no measurements are available other than those made on the basis of the original classification. Therefore, the Silver Standards will be used in chapter 6 to rate the outcomes of the final simulations.

6. Final conclusions

In the experiment that was presented in this article, experts judged whether a given classification line was suitable for a medical case which was described in a diagnosis report. For 107 of the 240 experiment reports (44.6%), at least one of

TABLE 8: silver standard III and expert rating

		expert rating			
		++	+/-	--	
original silver standard	++	90	39	15	144
	+/-	58	85	145	288
	--	23	64	569	656
		171	188	729	1088

Kappa	.401	(.024)
Spearman Correlation	.586	(.025)
agreement	.69	
disagreement	.035	

TABLE 9A: silver standard III and expert rating (Maastricht - excluding multiple and skin reports)

		expert rating			
		++	+/-	--	
original silver standard	++	27	17	12	56
	+/-	11	6	26	43
	--	7	12	124	143
		45	35	162	242

Kappa	.344	(.048)
Spearman Correlation	.566	(.053)
agreement	.65	
disagreement	.079	

TABLE 9B: silver standard III and expert rating (Helmond - excluding multiple and skin reports)

		expert rating			
		++	+/-	--	
original silver standard	++	46	11	2	59
	+/-	13	26	23	62
	--	3	21	98	121
		62	58	123	243

Kappa	.518	(.046)
Spearman Correlation	.718	(.038)
agreement	.70	
disagreement	.021	

the suggestions was rated to be a suitable classification. In 31.3% of the cases, none of the suggestions was useful.

The analysis of those suggestions that were identical to the original classification indicated that assigning classification lines to cases is prone to subjective interpretation. In a number of instances, experts judged the original classification as 'unacceptable'. This indicates that even if the nearest neighbor method is capable of finding conceptually very similar texts, it is hazardous to use it for unsupervised automatic classification until the quality of the corpus has reached a high level.

In a number of cases, a high text similarity did not guarantee a high expert rating. Reports on multiple examinations proved difficult for the nearest neighbor method to classify because of interactions between text segments. Reports on skin biopsies were often rated as partly useful or unuseful if the location of the skin did not match between report and classification. Judgment of cases was severe in general, partly because of characteristics of the experiment design.

The comparison of the suggested classification lines with the original classification line has fair predictive power with respect to the expert's rating. Three 'Silver Standard' calculations were introduced and assessed; they showed agreement with the experts in the order of .70. These Silver Standards use the original classification as a reference, although the experts rated some of the original classifications as 'only partly useful' or 'useless'.

The expert evaluation has pinpointed a number of weaker points in the similarity-based automatic classification of reports. Some of these points can be made stronger by applying additional procedures, and the improved 'Silver Standard' evaluation measure will be used for testing these procedures.

The results of this experiment underline that classification of pathology reports – manual or automatic – is not a trivial matter. For optimal use of large databases, the current variance that is due to personal preferences in coding should be dramatically reduced. Structured support in classification of medical cases is essential.

References

De Bruijn L.M., Verheijen E., Van Nes F.l., and Arends J.W.: Assigning SNOMED codes to natural language pathology reports. In: J. Brender et al (eds): Medical Informatics Europe '96. IOS press, Amsterdam 1996 pp 198-202

De Bruijn L.M., Hasman A., Verheijen E., Van Nes F.L., and Arends J.W.: Classification of diagnoses that are described in natural language. Accepted for publication in Int. J. of Technology Management, 1997

Findler N.V. and Van Leeuwen J.: A family of similarity measures between two strings. IEEE Trans. on Pattern Analysis and Machine Intelligence 1979 Vol PAMA-1 no 1 pp 116 - 118.

Friedman C.P. and Wyatt J.C.: Evaluation methods in medical informatics. Springer Verlag New York 1997.

Gantner G.E., Côté R. and Beckett R.S.: Systematized Nomenclature of Medicine Coding Manual, College of Americal Pathologists, 1979.

Hall P.A. and Lemoine N.R.: Comparison of manual data coding errors in two hospitals. J. clin. Pathol., 1986 Vol 39 pp 622-626

Landis J.R. and Koch G.G.: The measurement of observer agreement for categorical data. Biometrics 1979 Vol 33 pp159-174.

Wieggers J.G., De Bruijn L.M., Hasman A., and Arends J.W.: De gebruikersinterface voor de invoer van PALGA classificaties. (in Dutch) Accepted for publication in: Medisch Informatica Congres 1997.

CHAPTER · 6

SIMULATIONS REVISITED, GENERAL DISCUSSION, AND CONCLUSIONS

6 SIMULATIONS REVISITED, GENERAL DISCUSSION, AND CONCLUSIONS

1. Introduction

In the previous chapters, a method for automatic classification of pathology texts was presented. The general idea of this method is that an archive collection is searched for older reports that resemble as much as possible the new report that is to be classified. The 'nearest neighbor' of the new report was already classified (in order to be admitted to the archive collection), so if the resemblance is high enough, the classification terms of the nearest neighbor can be copied onto the new report. This method was tested in various simulations in order to test different versions of the similarity measure and of the vector space that forms the basis of the similarity computation. The simulations gave promising enough results to decide upon a large scale expert evaluation, which was presented in chapter 5 of this thesis.

From the experiment's results, it was concluded that the nearest neighbor method needed improvement in order to be useful. Reports on skin tissue were found vulnerable because near neighbors describing tissue with the same morphological deviations were often describing skin samples from different locations of the body. Likewise, the similarity computation provided terms on left and right sidedness with a correctness that was only little beyond chance. Multiple reports proved to give difficulties. Improvements are suggested and assessed in this chapter, which then concentrates on practical use issues, the general and final findings from the project, the further research on the nearest neighbor method that would be required or desirable, and final conclusions.

2. Simulations revisited.

The expert evaluation experiment gave, apart from direct observations on the accuracy of automatic classification, a large amount of data for validating the so-called 'Silver Standards'. These are evaluation figures that can objectively be computed. With such a 'Silver Standard' it is possible to predict the judgment of an expert, so that it is not necessary to bother the expert for judging new observations in large scale experiments. A large scale experiment was carried out: 6640 pathology reports were classified with the nearest neighbor method.

2.1 Basic simulations revisited.

experiment: Goal of this experiment is to establish how the nearest neighbor classification method performs with a broad range of reports, when evaluated with a Silver Standard.

material: 5000 reports from one archive (Maastricht), and 2500 reports from a second archive (Helmond) – derived as described in appendix 1 of this thesis – formed the two collections on which simulations were performed (see also appendix 1 of this thesis). Multiple reports, i.e. those that describe sections from different body locations and which are reported as such, were eliminated from the collections, leaving 4501 resp. 2139 singular reports for experimentation. The words were not transformed, nor corrected for spelling errors, typo's etc.; weight terms were added which were computed with the 'inverse document frequency' formula:

$w_{ij} = f_{ij} * \log (D_t / D_i)$;

the weight w_{ij} for word i in report j equals the number of occurrences of word i in report j (f_{ij}), multiplied with the log of the ratio between the total number of reports in the collection (D_t) and the number of reports in the collection that contain word i (D_i). This weight term is higher if a word is relatively rare, and drops if a word is more common (see also chapter 3 of this thesis).

retrieval method: for retrieving the nearest neighbors of a test report from the collection, the test report (r1) is compared with each report (r2) of the collection on the basis of a similarity function $T(r1, r2)$:

$T(r1, r2) = \sum_i w_{i,r1} * w_{i,r2}$,

which is normalised into $T_n(r1, r2)$ – lying between 0 and 1 – with:

$T_n(r1, r2) = T(r1, r2) / \sqrt{(T(r1, r1) * T(r2, r2))}$.

Per test report, k-Nearest Neighbors are retrieved such that the outcome results in five different classification lines.

simulation method: since the method has no 'training'-phase, the *search* collection may be used as *test* collection. Each of the reports in the collection serves as a test report and nearest neighbors are searched within that collection. Obviously, the test report is prohibited from retrieving *itself* as a near neighbor. This simulation method can be counted as a leave-one-out method.

evaluation method: the classification codes are known for all reports. They do not fulfill a role in *retrieval*, but are used to form the basis of the *evaluation*. A comparison of two code lines estimates the semantical similarity of cases; it is used to compare the original case with the nearest neighbors. If two code lines are identical, the cases are assumed to be very similar. Two notes are made:

◇ diagnostic coding is subject to personal variation: this was found in the expert evaluation (see chapter 5, section 4.2 of this thesis) and reported in literature (see chapter 2), so similar cases may be coded with slightly dissimilar codes. Thus, a dissimilarity between codes does not mean that the cases are totally different, or that retrieval has failed. In order to identify

those slightly dissimilar codes that still indicate identical cases, a similarity measure between SNOMED classification lines was constructed. Section 5 of chapter 5 (this thesis) is dedicated to the design and validation of the SNOMED similarity measure as a 'Silver Standard'.

◊ the classification lines may contain errors. Evidence was found in our expert evaluation (chapter 5 of this thesis) and was also reported in literature (Hall and Lamoine 1986). This will add a certain amount of noise to performance *and* evaluation. Still, the classification lines form the basis of the only possible independent estimation of semantical similarity and is therefore used. Error levels are assumed to be limited in this study.

Silver Standard III, which was validated with the data from the large scale expert evaluation (see chapter 5, section 5.3.4 of this thesis), was used. The semantical correspondence between two classification lines is estimated by calculating the string similarity for Topography codes in the two classification lines (Sim_T), as well as for the Morphology codes (Sim_M); an additional factor is P_{Skin}, which equals 1 if the report describes skin tissue, and 0 for other reports. The equation

$$\text{Rating Estimate} = .6 * Sim_M + .3 * Sim_T - .14 * P_{Skin} \quad ,$$

divided into one of the categories

++ essentially correct (Rating Estimate $> .75$)

+/- partly correct; can be made correct with a minor modification (Rating Estimate $\in [.53, .75]$); and

-- incorrect or useless. (Rating Estimate $< .53$)

was found to agree with expert's ratings in about 67% of the cases, and gave an opposite judgment in about 4% of the observations (see chapter 5 - 5.4). Note: since this experiment is restricted to singular reports, the factor for multiple reports is left out of the equation.

Results and discussion

Tables 1a and 1b present retrieval results for the first nearest neighbors, the best of the two nearest neighbors and up to the best result within the first five nearest neighbors. These figures describe the chance of finding a correct (or partly correct) classification within the first *k* suggestions.

The tables show that in about 25% of the cases the first nearest neighbor gives a good classification according to the Silver Standard. In about another 25% of the cases, the suggestion is partly good – can be made correct with minor editing. If one opts to select the best solution from the first five nearest neighbors, then a good suggestion is given in about 45% of the cases, and a good or partly good suggestion can be expected in about 75% of the cases.

For each test report, it was checked with the Silver Standard whether *any* of the classifications in the collection would suit the case (regardless of the textual similarity). This yielded the 'maximum'-figure in tables 1a and 1b: the number

of reports classifyable with the search collection. Results are subsequently related to these retrieval-maxima. For 2982 out of 4501 reports (66%) in the Maastricht collection, the classification lines in the collection included at least one correct classification – according to the SNOMED similarity measure. For 302 reports (4501-4199; 6.7%), none of the classifications would even be partly appropriate to the test report. In these cases, the 'haystack' is 'needleless' and the search mechanism cannot be reproached of failing to find suitable classifications. The Helmond collection contains a '++'-pendant for 64% (1379 / 2139) of the test reports; for 136 reports (6.4%) the haystack is needleless.

These results come out lower than the results in the other simulations that were run with the nearest neighbor method (Chapter 4). The previous evaluation measure was based on observations from a pilot experiment where the judge's task existed of rating the correspondence between report texts. The current evaluation measure was based on a large scale expert evaluation where experts judged how well classification lines suited report texts. The current measure is regarded to be more precise but also more severe. The absolute performance of the previous simulations should be replaced with the current observations. Considering the stability of relative performance, the conclusions that were made on comparisons between model varieties may remain unchanged.

TABLE 1A: correctness in *k* suggestions for Maastricht archive; with percentages of total and maximum. Total number of cases = 4501

k	++		+/-		--		++ or +/-	
1	1214	(27.0; 40.7)	1119	(24.9; 26.6)	2168	(48.2)	2333	(51.8; 55.6)
2	1622	(36.0; 54.4)	1264	(28.1; 30.1)	1615	(35.9)	2886	(64.1; 68.7)
3	1864	(41.4; 62.5)	1303	(28.9; 31.0)	1334	(29.6)	3176	(70.4; 75.6)
4	2005	(44.5; 67.2)	1330	(29.5; 31.7)	1166	(25.9)	3335	(74.1; 79.4)
5	2117	(47.0; 71.0)	1332	(29.6; 31.7)	1052	(23.4)	3449	(76.6; 82.1)
max	2982		4199		4501		4199	

TABLE 1B: correctness in *k* suggestions for Helmond archive; with percentages of total and maximum. Total number of cases = 2139

k	++		+/-		--		++ or +/-	
1	488	(22.8; 35.4)	613	(28.7; 30.6)	1038	(48.5)	1101	(51.5; 55.0)
2	710	(33.2; 51.5)	660	(30.8; 33.0)	769	(36.0)	1370	(64.0; 68.4)
3	822	(38.4; 59.6)	641	(30.0; 32.0)	676	(31.6)	1463	(68.4; 73.0)
4	899	(42.0; 65.2)	635	(29.7; 31.7)	605	(28.3)	1534	(71.7; 76.6)
5	958	(44.8; 69.5)	631	(29.5; 31.5)	550	(25.7)	1589	(74.3; 79.3)
max	1379		2003		2139		2003	

FIGURE 1A: Code term coverage for the first up to the eighth suggestion; Maastricht collection (4501 cases)

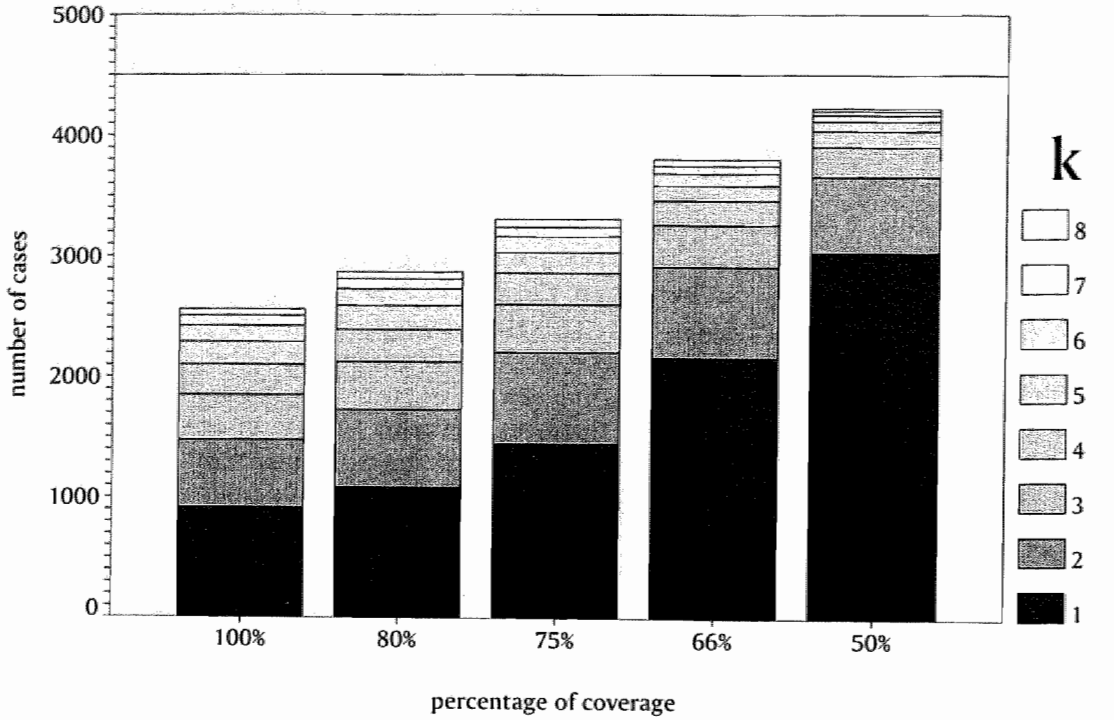
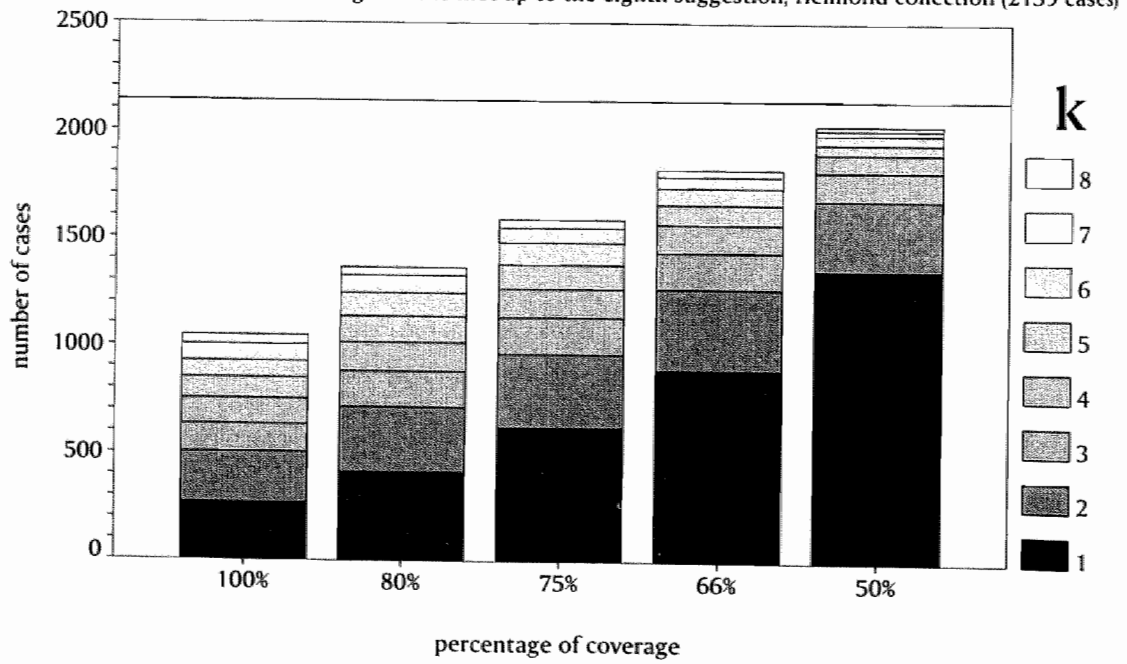


FIGURE 1B: Code term coverage for the first up to the eighth suggestion; Helmond collection (2139 cases)



The results in tables 1a and 1b show that, although useful classifications are given for the majority of cases, the number of incorrectly suggested classifications is so high that it puts question marks at the value of the method. Especially if it would be the easy cases where classification does well and if the method would be powerless with the more difficult cases.

The latter argument was explored by assuming a relation between the length of a classification line and the complexity of the case. Simple cases may be coded with only a few code terms whereas classification of more complex cases may require a larger number of code terms. The majority of the 6640 cases was originally classified with 3 (2132 cases), 4 (1911 cases) or 5 code terms (1199 cases). In these categories, the first suggestion was rated '- -' in about 47% of the cases, and this figure remained fairly constant for reports with longer classification lines. The first suggestion was correct ('++') in about 28% of the cases with 3, 4, or 5 code terms, for longer code lines this percentage lowered to 20%, thus trading off with the percentage of '+/-' rated cases. A similar pattern could be observed for the scores of the first five suggestions. These figures do not indicate a dramatic performance drop for cases with increased difficulty.

Systematic support would relieve the pathologist from a trivial though tiresome task, even for easy cases. The quality of classifications may rise: a pathologist will be given classification lines that were composed before by colleagues when they classified similar cases and this may invoke reduction of personal differences in classification. Furthermore, certain types of classification errors – those that are caused by laziness – *are* likely to occur especially in 'easy' cases. On the other hand, the risk that pathologists accept non-optimal classification lines because editing the classification line would be too much trouble, should not be neglected when developing the user interface.

Note that the method knows no difference between easy and difficult cases. Since the nearest neighbor method does not require any modelling, difficult cases are dealt with like any other case. The only difference lies in the rareness of cases – which is not unrelated to difficulty – because chances are that even the best match for a rare case differs on a significant number of characteristics.

Coverage: another measure for evaluating the usefulness of suggestions is based on *coverage*: the extent to which classification terms in the solutions cover the various classification terms in the original classification. In 1178 cases of 6640 (collections combined: 4501 reports from Maastricht and 2139 reports from Helmond), all terms in the original classification were contained in the first suggestion; in 3138 cases, all terms could be found in the set of terms from the first 5 suggestions. In 2046 and 4409 cases, 75% or more of the classification line could be reconstructed with the terms from the first resp. the first 5 suggestions. In 3047 and 5242 cases, the coverage of the first resp. the first 5 suggestions was more than two-third. Figure 1 summarizes these results.

The average number of different terms in the original classifications of singular reports was 3.8 (sd=1.3; range 1-11) for the Maastricht archive, and 4.5 (sd=1.3; range 1-11) for the Helmond collection. The average number of different classification terms in the first up to the fourth suggestion – this is where Figure 1 levels out – was 9.4 (sd=3.1; range 3-25) for the Maastricht collection and 12.2 (sd=4.2; range 4-28) for the Helmond collection. In practice, one would have to browse a set of on average 9.4 different classification terms to cover the whole of the original classification line for a Maastricht report, and have a chance of 47% that all terms are present in the set, and a chance of 77% that at most one word has to be added.

The simulations show that the nearest neighbor method is capable of retrieving cases from the archive that either propose relevant complete classifications for the new case, or give useful separate terms for classification line composition. The results are obtained by a method that is as generally applicable as possible because domain modelling is not incorporated to support the retrieval of cases. If the nearest neighbor classification method would be used on another domain, then the above results give a first estimate of the functional performance.

The number of cases where results are not satisfactory, brings about the necessity for refinement of the method. In the previous chapter, a number of enhancements was proposed. In this chapter, these enhancements are further developed and put to the test.

2.2 Enhancements: skin location

In the examination of skin tissue, microscopic observations on cell structure and morphologic change are reported. The exact location from which the biopsy or excision originates, is given in the report and will most often be included in the classification line. In many cases, the microscopic observations have little to do with the exact location but describe findings that are typical for skin tissue in general. With the nearest neighbor method, chances are that a similar case is retrieved that describes skin tissue indeed, with a similar morphologic change, but originating from a different location.

In order to adjust the suggested classification line to the exact location, an additional procedure was implemented. For each singular 'skin' report, the first phrase of the conclusion paragraph was extracted – starting from the first occurrence of the string *huid* (*skin*) and ending with the next occurrence of a punctuation mark (period, comma or colon) or either one of the strings *met* (*with*) and *waarin* (*in which*). This word string would typically look like '*huidbiopt linker onderbeen*' (*skin biopsy lower left leg*), '*huid slaap*' (*skin temple*) or '*huidschraapsel schouder rechts*' (*skin scraping right shoulder*). These word strings are stored together with the T(opography) terms of the classification line of the

report. Since left/right sidedness was dealt with in another routine (see section 2.3), these words were removed from the text strings, and the corresponding classification codes were deleted. The average number of words in the word strings was 2.2 (sd=1.5), the average number of 6-character Topograpy codes per report was 1.9 (sd=0.6). The word string extraction routine gave exceptionally long word strings (12 up to 28 words) for nine reports out of 1822. In these cases, the conclusion line was indeed atypical but this did not cause problems in processing. They were not changed nor eliminated.

When a new 'skin' report is being processed, the word string from the 'conclusion'-text can be derived analogously. The Nearest Neighbor method is again used to find the most similar 'skin'-phrases from the corpus of conclusion line word strings. The T-terms from the most similar 'skin' phrase then replace the T-terms that were originally suggested.

This procedure was separately tested in a simulation using all the singular 'skin'-reports. Only rarely, a single most similar phrase was found (18%): in most cases the identical phrase was found in several other reports, that then shared the position of 'nearest neighbor'. In 42% of the cases, the position of 'nearest neighbor' was pooled by ten collection items or more. This means that a decision must be made on *which* codes are to be suggested if SNOMED codes differ between pooled nearest neighbors. In this experiment, those codes are chosen that occurred the most often, and those that occurred in more than three out of four retrieved code sets.

In 1347 of the 1822 cases (74%), exactly identical word strings were present in the 'skin'-corpus. In 1371 of all cases (75%), the suggested classification codes matched exactly – even if the text did not match exactly (329 of 475 = 69%). In 81 cases (4%) the suggested classification codes differed more than 30% from the original classification codes.

In 349 reports from the Maastricht archive and in 35 reports from the Helmond archive, no specific codes were given for the exact skin location. In 91 (Maastricht) and 2 (Helmond) of these cases, the text did contain information on the exact skin location, so the actual classification codes did not exactly match the text. For optimal use of this routine, an accurate archive collection is therefore essential.

This leads to the conclusion that an additional routine can find in most cases the exact topography of a skin biopsy in the text, and assign appropriate classification codes. Such a routine can correct the insensitivity of the general nearest neighbor method to code the specific location of skin biopsies.

Enhancements: left/right sidedness 2.3

Left sidedness and right sidedness is an important feature of many reports. In 1451 reports (19.3%) of the total collection (7500 reports), the classification line contained explicit codes for left- or right sidedness, bilaterality or 'left/right

unknown'. In spite of its importance, the sidedness is described in the report by using the word 'left' or 'right' only once or twice. With the nearest neighbor method as implemented now, the key terms for sidedness 'drown' among the other words in the report. Whether the retrieved report agrees in sidedness may be a matter of mere chance.

A simple procedure was added to the retrieval system that tracked the sidedness of the report. The report was checked for a small number of words:

- ◊ 'links', 'linker' and 'li' to indicate left sidedness;
- ◊ 'rechts', 'rechter' and 're' to indicate right sidedness; and
- ◊ 'beiderzijds' and 'dubbelzijdig' to indicate bilaterality (a report is also 'bilateral' if both left sidedness and right sidedness is mentioned).

No additional procedure was used to intercept the cases that are explicitly labelled with 'sidedness unknown': it was not clear when this classification term is used and when not.

This procedure gave the results as given in table 2. In many of those cases in which the classification line described sidedness, the routine supported that description (95.0%). For 54 reports, the text contained phrases with both 'left' and 'right' but the classification line gave only left or right sidedness. Various reasons could be found:

- ◊ although the *tissue* came from one location, the patient history or the patient's complaints listed the other side as well (10 cases)
- ◊ tissue came from both locations but only one side was explicitly indicated in the classification line (16 cases)
- ◊ the string 're' also occurs in 're-resection' and in 're-excision' (even 're-re-re-excision' once) and misled the routine (9 cases). Left and right sidedness should be described in full and not abbreviated; this is what Silverberg (1996) recommends as N.A.I.T.D.L. (*no abbreviations in the diagnosis line*).
- ◊ in macroscopic examination and preparation, tissue may be marked on one side. This is indicated with, e.g., 'the left side of the material is marked green' (5 cases)

TABLE 2: agreement on left/right sidedness between report and suggested classification line

		classification line				
		left	right	both	unknown	total
report	left	564	3	0	463	1030
	right	0	582	0	527	1109
	both	27	27	195	292	541
	unknown	5	4	4	4807	4820
	total	596	616	199	6089	7500

◊ in 4 cases the classification line was wrong; in 5 cases the classification line was dubious;

The 3 cases in which a 'right side' was classified but the text indicated left sidedness, were not correctly classified. In some cases the classification gave a sidedness-code whereas the text contained no direct reference to the side. This was due to implicit reference ('vasa deferentia' and 'tonsils' imply bilaterality) or to enigmatic abbreviations ('ooglid OD', 'COPD', 'LOK' and 'LIMA-operatie').

The results show that this simple routine determines the correct sidedness in a large number of cases. It is not clear in which cases the sidedness is explicitly coded and in which cases such a code is not given. Those cases where the routine did not concord with the classification line often indicated incorrect or dubious coding, or untidy reporting.

Enhancements: key terms 2.4

As with left/right sidedness, there are other concepts that are reported in a single word but that are important for classification. Notably these are: degree of dysplasia or bacterial colonisation, resection edge free or not free, additional testing or staining positive or negative. For each of these concepts, ad-hoc routines can be constructed to intercept erroneous code suggestions. Such routines were not implemented as yet. Still, a few words on the subject are spent in the next paragraph.

Cases such as mentioned are not rare: dysplasia is coded in 215 reports, with 233 classification codes giving light dysplasia (77), modest dysplasia (91), severe dysplasia (45) and dysplasia without a specified degree (20). For 384 reports, the state of the resection edge is coded (407 classification codes) describing a resection edge that is free (264) or not free (143). For cases such as these, errors in the classification line may be intercepted with additional routines.

Malignancy is seldom described in a single word. If the word '*maligniteit*' (*malignancy*) occurs in the text, its context negates its meaning in 4371 of 4470 occasions. If the word '*maligne*' occurs, it is often used in the 'clinical questions' paragraph as a query (111 times out of 343 occurrences of the word). Therefore, the occurrence of 'maligniteit' or 'maligne' certainly does not imply that malignancy *was* observed in the material. If the system would add a 'no malignancy' code term to the report whenever the word 'maligniteit' is encountered, it would perform nearly perfect, but this would give a very undesirable effect in those cases where malignancy *is* present. A final solution to this question was not yet found.

For a number of key terms, interception routines may be devised. In the current study, this was not yet undertaken. Terms such as *malignancy* prove more difficult to intercept and demand deeper modelling of negation phrases, and may require subtle disambiguation. In an interactive system, the routine

may consist of the system detecting a key term in the report and prompting the user to accept one of the possible alternative code terms.

2.5 Enhancements: multiple reports

Multiple reports are those reports that describe different sections, often with different diagnoses. Sections are referred to by Roman numerals, occasionally subsectioned with letters (e.g. *IIb*). Sometimes the sections are highly similar – for example naevi that were removed from different locations of the body, or different regions of a single stomach. It is not always clear when different sections are to be described in the same report and when separate reports are made for different material.

When the textual similarity is applied as in this study, the context is removed so that it is not retraceable anymore to which part of the report a word belongs. For singular reports this is no major objection, since an occurring concept will relate to the single subject that is under consideration. In a multiple report, 'round nuclei' can refer to the 'stomach' part of the report, to the 'duodenum' part of the report, or to both.

Multiple reports take an important part in reporting. In the collection of 7500 reports, there were 860 multiple reports: 715 double reports, 121 triple reports and 24 quadruple reports. They cover a diversity of topics, with a number of typical combinations:

- ◇ 'cervix/uterus' reports (260 cases),
- ◇ intestine sections (284 cases),
- ◇ 'skin' reports (130 cases),
- ◇ reports on breast tissue and lymph glands (64 cases), and
- ◇ 'urology' reports (34).

A solution for handling multiple reports can be found in separating the texts fragments that treat the different sections. This can be done at the time of reporting – by using structured text entry – or afterwards – by splitting the sections and regrouping them per section. The latter method can be guided by the subsection headings that are often used throughout the report and that function as cross references between paragraphs. The method is hindered by references to previous parts of the text, e.g. 'IIb: the same structures are seen as in section IIa' or grouped observations, e.g. 'sections II to IV: ...'. It may be worthwhile to systematically guide the reporter in using a clear paragraph structure when a multiple case is under examination.

In automatic classification, the subsections of the multiple report can be grouped so that they describe a distinct subject, and then subjected to Nearest Neighbor classification. If this is sequentially done for the several sections, a classification line can be found for each separate section. Methods for handling multiple reports were not worked out in this study, so no results are available for evaluation.

Simulations revisited after enhancements 2.6

A further simulation was run, but now incorporating two of the presented enhancements: checking the exact location of skin reports, and checking left/right sidedness of source tissue.

The procedure of this simulation was as follows: the most similar reports were retrieved in exactly the same manner as described in section 2.1 of this chapter. They provided the candidate classification lines. If appropriate, the enhancement routine for 'skin' reports (section 2.2) was executed, and the suggestions for the classification lines were adapted. Then, if appropriate, the enhancement routine for left/right sidedness (section 2.3) was (also) executed, and the suggestions for the classification lines were (again) adapted. The results of these simulations are listed in tables 3a and 3b .

A comparison with the results from the original simulation proved that the enhancements have improved the quality of the suggestions. This improvement was statistically significant for the set of those reports that were subjected to additional processing, as well as taken over the entire set of test material (paired t-tests; $p<0.01$). A comparison of tables 3a and 3b with tables 1a and 1b shows that the improvement for useful classifications ('++' or '+/-') remains

TABLE 3A: correctness in *k* suggestions for Maastricht archive after enhancements; with percentages of total and maximum. Total number of cases = 4501

k	++		+/-		--		++ or +/-	
1	1386	(30.8; 40.6)	1024	(22.8; 23.7)	2091	(46.5)	2410	(53.5; 55.7)
2	1842	(40.9; 53.9)	1120	(24.9; 25.9)	1539	(34.2)	2962	(65.8; 68.5)
3	2108	(46.8; 61.7)	1121	(24.9; 25.9)	1272	(28.3)	3229	(71.7; 74.7)
4	2271	(50.5; 66.5)	1120	(24.9; 25.9)	1110	(24.7)	3391	(75.3; 78.4)
5	2394	(53.2; 70.1)	1106	(24.6; 25.6)	1001	(22.2)	3500	(78.2; 81.0)
max	3416		4323		4501		4323	

TABLE 3B: correctness in *k* suggestions for Helmond archive after enhancements; with percentages of total and maximum. Total number of cases = 2139

k	++		+/-		--		++ or +/-	
1	522	(24.4; 36.6)	617	(28.8; 30.1)	1000	(46.8)	1139	(53.2; 55.5)
2	745	(34.8; 52.2)	657	(30.7; 32.0)	737	(34.5)	1402	(65.5; 68.3)
3	853	(39.9; 59.8)	646	(30.2; 31.5)	640	(29.9)	1499	(70.1; 73.1)
4	931	(43.5; 65.2)	638	(29.8; 31.1)	570	(26.6)	1569	(73.4; 76.5)
5	985	(46.0; 67.1)	635	(29.7; 30.9)	519	(24.3)	1620	(75.7; 78.9)
max	1427		2052		2139		2052	

constant irrespective of the number of suggestions that is considered: better performance for about 60 cases from the Maastricht collection, and about 36 cases from the Helmond collection. The value of the enhancement routine increases with the number of suggestions for reports from Maastricht in the '++' category: for $k=1$, enhancements improve total performance with 3.8%, for $k=5$ this is 6.2%. For reports from Helmond, this figure lies steadily around 1.5%. Since the maximum obtainable performance (given the archive collection) has also risen with the enhancement routines, the *relative performance* is incidentally lowered.

The enhancements have led to improvement in the performance. For those cases where an enhancement routine was run, the result was almost invariably successful: the retrieved items were more relevant after adaptation or remained equally relevant if the retrieval was successful in the first place. The final classification performance has still not reached a high level. Chances are about 53% that the first classification is either correct or is useful as a basis for classification, and about 78% that a useful suggestion can be found within the first five alternatives. Relative performance lies at about 70% for a *good* classification within the first five alternatives, and about 80% for a *useful* classification within the first five alternatives.

The experiments have shown that the nearest neighbor classification method is capable of extracting from an archive collection a small number of reports with a high chance of them being relevant. The high chance is not yet high enough to consider autonomous classification. Significant improvement of the method through changing parameters is not to be expected: making word variations uniform, adapting term weight factors, and using other similarity measures have been shown to give a small improvement where a larger step is needed (De Bruijn et al. 1997). Additional operations for intercepting specific groups of cases can also give improvement, but only for those cases where an interception rule is available, is successful and where the original effort has failed. The merit of a larger archive collection for functional performance is interesting: the increase of chance for very similar reports being present in an archive collection almost certainly outweighs the concomitant 'false hits'. A larger archive collection however also makes the technical performance (notably speed and memory requirements) more important. Finally, the key to success depends for a great part on the design of the user interface that supports the composition of classification lines: coverage figures imply that the nearest neighbor retrieval method can set the context for a 'smart' user interface in an easy and dependable way.

3. General discussion

A nearest neighbor method for text classification was proposed in chapter 3, further explored in chapter 4, evaluated by experts (chapter 5), and finally tested in a large-scale simulation (current chapter). The method showed a capability of relating new reports to archive reports that describe similar cases. This resulted in a *good or partly good* classification being given within the first five suggestions in about 75% of the 6640 test reports. In about 50% of the cases, one of the first five suggestions was a *good* classification. The results indicate that the nearest neighbor method is capable of aligning natural language reports such that neighboring reports describe cases that are similar.

The precision of the text similarity measure does not correspond with the desired level of detail in the classifications: retrieved reports are similar, but not similar enough for their classification to be perfect. Slight differences on a number of aspects (or less slight on one aspect) made the experts judge the classification line as 'poor'. In these cases, changes should be made to the classification in order to make it tailor-fit. It is possible to define a threshold on the text similarity measure that guarantees good resulting classifications. The higher precision will, however, be accompanied with a high rejection rate so that cases remain unclassified.

In literature, methods based on syntactic or semantic analysis of texts were reported to perform successfully in 54% to 71% of the cases (Sager 1982, Briegl et al. 1994, Spyns and De Moor 1996). Method and test material in these studies were short pieces of text and restricted to limited domains: discharge summaries and head cancer reports. The nearest neighbor method, applied to a wide range of full pathology reports, gave flawless performance in a lower percentage of cases. When the first few suggestions are considered, or when partly good classifications are counted, performance of the nearest neighbor method lies within the range that was found in literature for other systems.

These observations imply that the nearest neighbor method alone is not strong enough. It *is* capable of taking a first step, where the gap towards perfect classification line is taken by another method or a person.

In a sequentially designed system that uses text similarity as a first step, and then performs precise semantic or syntactic analysis, the intermediary results may feed the algorithms of the second step in order to perform more precisely or faster. The investment in human effort for developing knowledge based techniques may be reduced with such an architecture. Croft (1993) advocates the design of hybrid systems in which statistical and knowledge based elements promise more effective systems.

A system may use the nearest neighbor method to interactively support the user in classification. Potentially good classification lines can be suggested, and

alternative terms can be handed to adjust the classification line to the specific characteristics of the case. This setting agrees with the original goal of the project: providing additional functionality to a reporting system in which automatic speech recognition is used for data entry. In that scenario, a report is dictated and immediately processed by the speech recogniser, and classification lines are suggested for that case. After the examination session, the pathologist verifies the report text and accepts (one of) the classification line(s), possibly after altering it. In such a scenario, integration with the daily working routine is needed; furthermore, interactive use requires high system speed and a careful design of the user interface. Sections 3.1 to 3.3 will discuss these issues further.

An interactive system has a number of merits over the current situation. In the first place, SNOMED classification of pathology cases is a task that is burdensome and error-prone, and also sensitive for subjective variation. This was reported in literature (see Chapter 2 Section 6), and our expert evaluation showed the same. It is easier for a pathologist to assess classification lines that are presented than to think up new classification lines. Presented classification lines can come from collegues, so classification may become more standard due to the implicit communication that takes place between pathologists. For the selection of additional terms, a small subset of the SNOMED thesaurus – generated on the basis of the report – will make searching easier. The collection of reports (and the classification lines attached) may initially contain errors and subjective colouring, but this can decrease after a certain offset-time if an algoritm is used for automatically refreshing the archive.

An experiment that was run with a prototype of the classification system, indicated that pathologists changed their classification style under influence of the system (Wieggers et al. 1997). They used slightly more terms to classify cases when using the system compared with manual coding, but this difference was not significant (4 pathologists, 30 reports, 4.73 terms per report (with computer), 4.60 (manually) – $p>.05$). The assigned classifications were more consistent between subjects when the computer was used than with manual classification ($p=.01$). Computer-supported classification took more time than manual classification, but this seemed to be due to the novelty of the method.

3.1 Integration with daily routine

In the current daily working routine, material is subjected to gross examination, processed into slides and then examined microscopically. The findings are reported through dictation which is typed out by secretaries. If the macroscopic findings are reported and typed out right after examination, the way in which it is done in most laboratories, the time that is required for preparation of the microscopy slides allows that part of the report is already electronically available when the microscopic observations are made and reported.

The classification of a case is usually dictated with the rest of the report, directly after the 'conclusions' section of the natural language report. With the current state of technology, the full report is not electronically available for automatic processing. Automatic support for classification must then be done (1) on the first half of the text (paragraphs on clinical data, query and macroscopic observations) that *is* available, or (2) on the full report after the dictation is typed out, or (3) on the full report if entered via another (less time-consuming) method than typing.

ad (1): The same classification method can be used for smaller portions of text. The first half of a report is sometimes only a few words long, but at times can take up to some tens of lines. The macroscopic characteristics predict the microscopic observations to some extent and restrict the possible relevant classification terms to a large extent. The *k*-nearest neighbor method would need a larger value of *k* because it is likely to have a lower precision, but may still give a significant decrease in search time when assisting the pathologist. Further investigation on this matter would be needed before such a classification aid can be put to work.

ad (2): In this case, the classification is added to the case when the report is verified – which may be a few hours or even days after the actual examination. Whether the quality of the classification suffers from the time lag or not, is a question that cannot be readily answered.

ad (3): At the start of our project, the possibility of diagnosis reporting with speech recognition technology was considered. It was found that the state of the art on speech recognition technology should progress before it could be applied successfully. However, in anticipation of this progress, direct automatic classification of completely dictated reports was tackled. Additional functions, such as classification support, would make acceptance of speech technology by pathologists easier.

Until speech recognition technology has improved, strategies (1) or (2) may be opted for.

3.2 Technical performance: speed

A system such as presented here – one that relies on searching a large collection of items – may pass results too slowly to be of practical value. In initial experiments, the time to retrieve relevant items from the collection amounted to about two minutes. This is clearly too long for close interactive use. Although the technical performance was of less interest in the present studies, it is useful to address the issue of speed for the sake of practical credibility.

Computer technology progresses, so the system's speed will go up due to increased computation speed, decreased memory access time and increased memory storage capacity. The initial simulations were performed on a PC

configuration with a 486 DX 33 MHz processor and with 8 MB RAM. Later upgrading to a Pentium 100 MHz 16 MB RAM system gave faster performance (no precise measurements were made). However, a more systematic improvement is needed, because above observations are based on the current state of the collection. It was shown that a larger archive collection would be needed in order to better cover the domain of pathology and to improve classification of cases. A larger collection slows down retrieval.

First, it is not sensible to add every next report to the collection, because then the collection would clutter with large numbers of routine reports. A collection can be kept trim if older reports 'decay' and if cases that have a sufficient number of similar counterparts in the collection are not included. Actual boundaries for 'age' and 'routine similarity' would have to be set.

Second, the current retrieval approach was not very sophisticated: every item in the collection was considered as a candidate for the set of k nearest neighbors. With alternative architectures for the collection, a stepwise search strategy can be adopted and search time can be decimated. Either an inverted tables architecture, or a cluster organisation of the collection, may enable stepwise searching.

One option for faster searching, is by using additional *inverted tables*. These tables indicate which items from the collection have at least some terms in common with the new text. Those items that have no or only a few terms in common, are discarded right away. The potential archive items are subjected to full comparison with the new text in order to determine the nearest neighbors. The inverted tables add redundancy to the collection, but this increase in memory storage requirements will trade off with a reduction in retrieval time. Note that with the advent of CD-R technology the cost of additional data storage is very low. The inverted table technology was not yet brought in practice in this project.

A different option for faster searching is *clustering* the items in the collection. This was tried in the project with a two-level cluster architecture: the term vector of a new text is compared to a set of centroid vectors that direct the search mechanism to (the) cluster(s) that are likely to contain candidates for the final k nearest neighbors. The clusters for the pathology collection could follow the division between the medical domains that are relatively independent, e.g. skin tissue, bone material, and intestinal samples. Time is gained because a large proportion of the collection is left outside consideration at an early stage of the search.

There are several cluster architectures possible, and numerous ways to assign items to the structure. Although a detailed study on the field of clustering falls outside the focus of this thesis, one method was explored to serve as an illustration on the feasibility of clustering. The following procedural approach was used to construct the set of clusters.

The similarity between any pair of items (reports) from the collection is determined (or this is done for a large subset of items), and these item-pairs are sorted in descending order of their similarity figure. Thus, the first line contains the pair of reports in the collection that are the most similar to each other. This ordered list is processed line by line. The reports on the first line form the core of the first cluster. Then the next line is read, and one of three rules applies:

1. none of the items of the pair was encountered before in the list; then the pair forms the core of a new cluster.
2. one report of the pair was encountered before; then the other report is added to the cluster that already contained the first report.
3. both reports were encountered before, and they were previously assigned to
   - 3.1 the same cluster: this corroborates the validity of that cluster, and then nothing needs to be done.
   - 3.2 different clusters: these clusters are now candidates for merger and this is noted as evidence. Clusters merge if eventually enough evidence is found ('enough' can be defined in terms of 'number of cross-cluster pairs' and the number of items in either pair – parameters that can be tuned).

Finally, a threshold on similarity scores stops the clustering process. Stopping too soon would leave too many items 'unclustered'; when stopping too late, clusters would keep merging. Apart from some large clusters, a number of very small clusters was formed, and some 'unique cases' were not assigned to any cluster at all. The small clusters and the 'unique cases' were put in a 'leftover-bin' cluster. The 'evidence for merger' was monitored manually, which led to merging a number of clusters.

With the 7500-report archive, initial clustering yielded about 420 clusters, after merger 148 clusters plus one 'leftover' cluster remained. This last cluster contained 983 reports from the site-1 archive (19.7%) and 258 reports from the second location (10.3%). Thus, a total of 83% of the items could be assigned to a 'real' cluster.

Each cluster, except for the 'leftover' cluster, is represented by a 'centroid', a vector or artificial item. This item may be constructed such that the similarity comparison of a new report with the centroid vector results in a figure that equals the average similarity between the new report and each of the items in the collection. Alternatively, a centroid vector can be composed of those features that are typical for the items in that cluster – i.e., those features that occur many times within the cluster and only very seldom outside the cluster.

With such a clustered archive, retrieval takes place as follows: the report to be classified is compared with each of the centroid vectors, resulting in a list of clusters that are likely to contain relevant items. The member-reports in these clusters (or in this single cluster) are assessed. The leftover cluster is consulted

FIGURE 2A: Patient data screen of prototype system

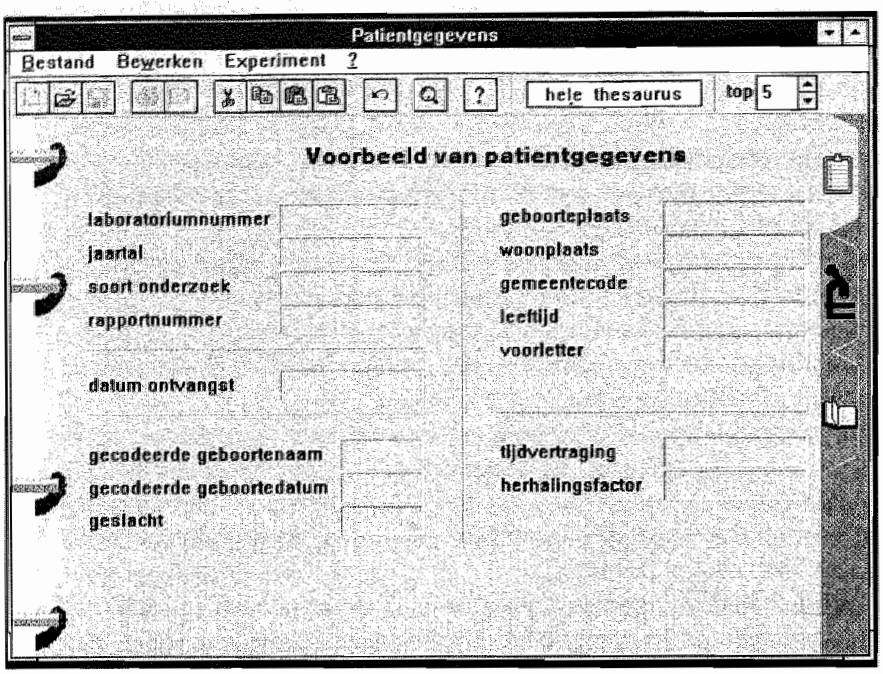
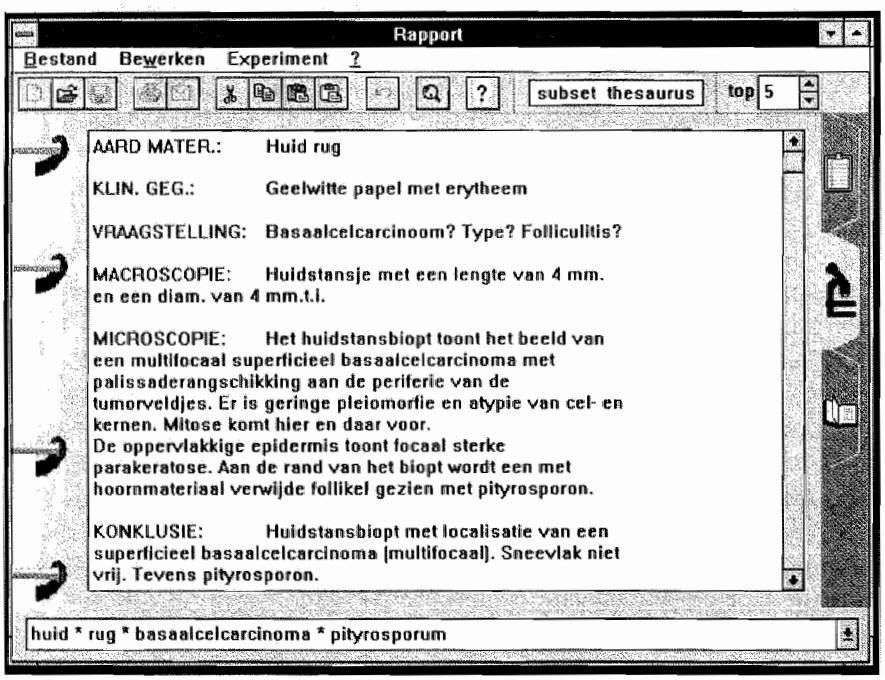


FIGURE 2B: Free text screen of prototype system



if centroid comparison indicated that it was useless to access any of the clusters, or if results from other clusters are still dubious.

The retrieval time will vary depending on which clusters are accessed, and on the setting of the parameters that determine the consultation of a next cluster or the leftover cluster. With our collection, retrieval time for one trial was reduced from about 55 seconds for the unclustered archive to about 1.2 seconds for the clustered archive (measured over 100 trials). Again, exact validation of clustering techniques lies outside the scope of this thesis.

User Interface Design 3.3

The user interface – that part of the system with which the human comes into contact physically, perceptually or cognitively (Maddix 1990) – determines to a great extent whether or not automatic support in classification becomes a usable function for the pathologist. Proposals for a user interface to the classification system are now in development in our department (Wieggers et al. 1997), and a prototype was tested at the time of this thesis being written. The following paragraphs illustrate what a user interface may look like, and give some design considerations with regards to the characteristics of the pathologist and the task of classifying cases.

In the design of the user interface we aimed at building on 'recognition' rather than on 'recall'. The human mind proves to perform better in recognition tasks than in recall tasks (cf. Baddeley 1976). A user interface is recognition based when functions are available through menu selection or button clicking rather than through use of key combinations, command word entering or parameter input. Recognition gives greater ease of use and causes less errors than recall. Dialogue in the user interface relies on menu selection rather than data entry with the keyboard, both for access to functions and for selection of data items (the SNOMED classification terms, notably). Input errors and interaction time can thus be minimized (cf. Norman 1991). Buttons on a button bar are available for a number of functions which can also be addressed through text based menus and function key shortcuts. This also supports powerful and fast interaction (cf. Ellis et al. 1995).

The data of the report is functionally divided for monitoring or editing on three pages or screens: (1) patient and administrative data, (2) the report text, and (3) the electronic thesaurus and classification line editor. Figure 2 shows the layout of these screens in the system's prototype.

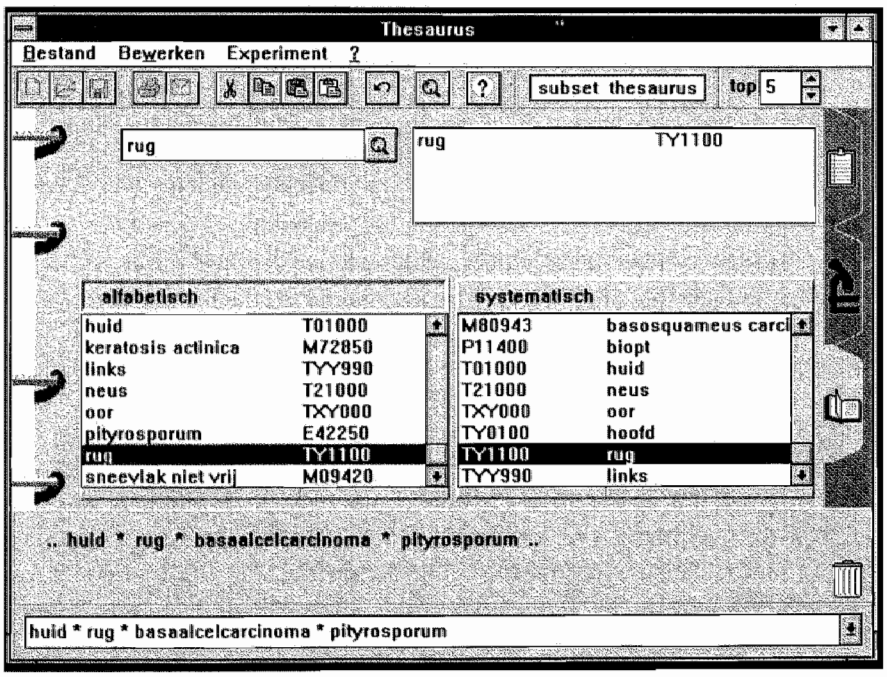
On the first screen (figure 2a), several details are available, e.g. the patient's age, gender, place and date of birth, the laboratory code, the report type and possibly the repeating factor of the examination. The second screen (figure 2b) exists mainly of a text field where the report is edited. Text editing functions are added here. After completion of the report, a classification line can be requested which is directly supplied through the Nearest Neighbor classification

method. The number of different suggestions that occur in the (drop down-) list can be set by the user. The user can select that classification that is best suited for the case or that gives the best starting point for tailoring a classification line.

A separate page is dedicated to editing the classification line (figure 2c). The lower half of the page presents the classification line itself, which can be edited with typing sequences, using menu options or dragging and dropping. The upper half of the screen displays an electronic version of the SNOMED thesaurus. Figure 2c shows that the thesaurus occurs in two panes: one in alphabetical order of the classification term, and the other systematically ordered through the classification code. The two panes are connected and will highlight the same item but in its two different 'environments'. The systematically ordered pane will be useful for searching classification terms that are related in meaning. When the user selects a term in the classification line, the thesaurus is set to that term. The thesaurus can also be used to extract terms for inclusion in the classification line. A search function is provided to find classification terms that contain a given character string.

When using the thesaurus, one can access the entire SNOMED thesaurus, but one may also opt for using only a subset of terms – those that are likely to occur in the classification line given the reporting text that is under

FIGURE 2C: Patient data screen of prototype system



consideration. The subset is filled with about 10 to 100 terms by collecting those terms that occur in the classification lines of those 50 reports that are textually most similar to the new report text.

In an experiment, four pathologists composed classification lines for 30 reports, doing 15 'by hand' – i.e. using the classification handbook – and doing 15 with computer support.

The subjects used the first nearest neighbor as a basis for classification in 82% of the cases. This percentage indicates that retrieval gives fruitful results. One of the next nearest neighbors was selected in 18% of the cases. The selected line was readily accepted (20 times, 33%) or changed with keyboard entry or mouse interactions. In the experiment, classification by hand was significantly faster than computer-aided classification. This may well be attributed to the novelty of the computer system: in the computer condition, the last third of the test set was handled faster by the subjects than the first third of the report set. The questionnaire revealed that the pathologists found the system easy to understand and to use. They thought that the suggested classifications offered them terms outside the 'vocabulary' they normally used.

Computer support resulted in more uniform classification lines than manual classification. If two pathologists classify the same report, the average correspondence between the classification lines is 0.78 (sd = .15), calculated with the string comparison figure (cf. chapter 5 of this thesis). When both pathologists had computer assistance, the comparison figure was .87 (sd = .11), when no computer assistance was available, this figure was .77 (sd = .11). Pairwise analysis of this data indicated a significant difference ($p=0.013$) (Wieggers et al. 1997).

It was concluded from this study that pathologists do not optimally correspond when classifying the same cases; computerised support can improve correspondence. Subjects were satisfied with usability, as with the classifications they wrote with the system. They think that they will be able to use the system faster after some use. The first suggestion formed a good start for the final classification, but alterations are often made.

4. Further research

As was indicated before, some important aspects of automatic classification have remained outside the scope of this thesis, partly due to temporal circumstances. Topics for further research include:

◇ *cluster organisation.* In the current study, clustering was done for practical reasons and without extensive validation. Clustering may add to better technical performance *and* lead to an increase in functional performance. The

number of 'false hits' may decrease if reports with a treacherously high textual similarity are left out of consideration in an early stage when their cluster is discarded in the first step. Precision could further increase if multiple cluster-levels were used: a term may be powerful to discriminate *between* clusters, but it has little discriminative power in the following stage, thus *within* the cluster. For instance, within a cluster of all 'colon' reports, the word 'mucosa' is of little value to discriminate between items. Recomputation of weight factors for the various levels in the cluster hierarchy may lead to more relevant retrieval results.

◊ *collection size.* In chapter 4, the issue of collection size was shortly discussed. In this chapter, the 'boundaries' of the collection became visible; a greater collection size may increase the chance that textually and semantically more similar items are present for a new report. We would have liked to extend the Maastricht archive with an additional portion of texts and repeat the simulations.

◊ *other collections.* Considering the differences that were found between the two collections that were used in this study, additional simulations on a third and perhaps a fourth collection would be worthwhile.

◊ *alternative weighting schemes.* The inverse document weighting scheme has been found to give good results. Nevertheless, additional simulations in order to assess alternative weighting schemes could be valuable, mainly because all side conditions are now satisfied for doing such simulations. Notably, the collections are present in a formatted manner, and an evaluation measure gives the occasion for easy and consistent judging of the results.

◊ *further improvement.* In section 2 of this chapter, some additional routines were presented and discussed. An increase in performance in automatic classification may be reached by including more routines. Notably, a routine that splits multiple reports would be of great value. The inclusion of linguistically based or probability based routines could aid in linking the Nearest Neighbor method with alternative programs in automatic classification, especially the knowledge based approach that takes a central position in a number of international projects (cf. section 7 of chapter 2, this thesis).

◊ *usability testing.* As was argumented before, the user interface is the essential factor for successful practical implementation because with the current performance, automatic classification needs to be supervised. The program that searches classification lines for free report texts has been completed with a user interface so that suggested classification lines can be edited: classification terms can be removed and terms from the on-line thesaurus can be pasted in the suggested line. This user interface is evaluated in an experiment that is in conduct at the moment.

◊ *alternative applications.* The nearest neighbor method may be usable for classifying other texts than pathology reports. The current study was designed

to be as context independent as possible, so that the results may be transposable upon experimentation with other corpori. But there may be tasks other than text classification where the method can be useful within pathology. Especially since the case context is known in advance (because clinical observations, examination query and the kind of material is described on the request form), the nearest neighbor method can be put to use in searching similar cases so that help can be given in a context sensitive manner, and reporting can be supported.

A preliminary exploration showed that the vocabulary of the second half of the report depends on the words that are used in the first half of the report: reports that look alike on the first half, are likely to look alike on the second half. This can be used to support automatic speech recognition. The second half of a report, the 'microscopy' and 'conclusions' paragraphs, is often written on the last day of the case's route through the laboratory so automatic speech recognition can only reduce turnover time when applied here. The first paragraphs of a test report (repeated with 490 reports) was used to retrieve similar reports with the Nearest Neighbor method, and the vocabulary of the second half was predicted on the basis of these past reports. The tailor-made vocabulary, which varied around 850 different words, covered 93% of the words (or 91% of the *different* words). A static set of 1000 words (optimally chosen) covered 89% of the words (85% of the different words). The prediction of narrative on the basis of past reports could reduce the error rate of a speech recognition system when used for report dictation.

5. Final conclusions

Classification of pathology reports is on the one hand an issue of which the importance is widely acknowledged. It forms on the other hand a task where inconsistency and subjectivity cause undesired variation in the product, or where its results are contaminated with pure error. The wide range of pathology diagnostics made the coding terminology grow over the years to a size of 15,000 to 25,000 terms, a scale that requires an encoder to consult the dictionary when classifying cases. Laziness and reliance on memory cause errors even in relatively simple cases, as was discussed in the literature review of this thesis (chapter 2). This is why automatic classification has been coined as a solution for the detected problems.

The new approach for pathology report classification that was introduced in chapter 3 of this thesis, is founded on principles from Pattern Recognition and Information Retrieval. A report is classified with the same classification line as the most similar text from a collection of earlier annotated texts. This method

is named 'Nearest Neighbor classification'. Where the principles of the Nearest Neighbor method are simple, the further requirements for successful implementation are more difficult.

In an experiment, a number of classifications was rated by experts in order to evaluate the retrieval method. Their judgements indicated that in 21% of the experiment reports, the first nearest neighbor gave a good classification, for about 45% of the test reports, one of the first five different suggestions was good. The experiment confirmed that in practice, errors and subjective variation contaminate the classification lines. Furthermore, the far from perfect values of Kappa, agreement and reliability supported the claim that the use of the standardized classification language is not as easy and consistent as usually assumed.

The wide range of texts on which the Nearest Neighbor method can be applied, brings along difficulties in evaluating the retrieval performance. An evaluation that does justice to the wide range of pathology reporting, requires a large scale setup. Only domain experts could give dependable evaluation results, but such resources are scarce. An alternative was sought in a Silver Standard evaluation metric, that was computed by comparing the original classification line of a case (these were available for the set of test reports) with suggestions for classification lines if the report is subjected to a Nearest Neighbor classification. Several Silver Standards were proposed and validated with data from a large scale expert evaluation. One of these Silver Standards, one that corresponded with the experts' ratings in about 67% of the cases and that gave an opposite judgment in about 4% of the cases, was used to evaluate a large scale simulation on 6640 pathology texts.

For about 66% of the cases, the collection contained a good classification. A good classification could be found within the first five nearest neighbors in about 71% of these cases. In about 76% of all cases, the first five nearest neighbors contained a good classification or a partly good classification. The outcome agrees well with the results of the expert evaluation.

If the first five suggestions are combined into a set of terms, then this set entirely covers the original classification in 47% of the cases, a coverage of >75% is reached in 66% of the cases.

Enhancements were developed to buttress detected weak spots (see chapter 5): the exact location of the source of skin biopsies, and left/right sidedness. Evaluation of these enhancements gave satisfying results. Additions concerning key terms such as 'malignancy' and 'degree of dysplasia', and concerning multiple reports were proposed but not further tested. Simulations were repeated after incorporation of the enhancements into the retrieval system. Absolute performance improved: about 53% of the first classification was rated to be at least partly good, and in about 78% of the cases a useful suggestion

could be found within the first five alternatives. The 'retrieval' ceiling also raised, so that relative performance remained the same.

The methods that make part of the approach are simple and elegant but gave promising results. For those collections of pathology reports on which the methods were applied, the textual similarity between report texts showed to result in retrieval of relevant archive reports. It also appeared that exact correspondence on detail was more difficult, so that textually highly similar but in meaning slightly deviant narrative emerged in the top of the retrieved items.

So, although the nearest neighbor method proved to be able to take a good step towards automatic classification, autonomous classification cannot be considered yet. The method *can* be used to give interactive, automatic support to the pathologist. Implementation of classification support has been completed in a prototype system that can be integrated with the daily working routine. Technical performance is sufficiently high: the average time to retrieve suggestions out of the collection of 7,500 reports was 1.2s. Since it was acknowledged from the experiments that suboptimal suggestions can be expected, extensive editing functions were incorporated in a 'direct manipulation' user interface. An electronically accessible thesaurus is included in the user interface.

Perfect performance is essential in classification of pathology reports. Automatic classification is possible but the gap with perfection is still too high to consider autonomous automatic classification. The nearest neighbor method did prove powerful to support a pathologist in writing classifications. In all, the computer-assisted pathologist probably gives the best possibility for feeding the national archive with good classifications.
Classifications that are correct, complete, consistent and concise.

References

Baddeley A.D.: The psychology of memory. Harper, New York 1976.

Brigl B., Mieth M., Haux R., and Glück E.: The LBI method for automated indexing of dianoses by using SNOMED: Part 2: Evaluation. Int. J. Biomed. Comput. 1995 Vol 38 pp 101-108

Croft W.B.: Knowledge-based and statistical approaches to text retrieval. IEEE Expert 1993 p 8-12

De Bruijn L.M., Hasman A., Verheijen E., Van Nes F.L., and Arends J.W.: Classfication of diagnoses that are described in natural language. Accepted for publication in Int. J. of Technology Management, 1997.

Ellis J., Tran C., Ryoo J., and Shneiderman B.: Buttons vs. menus: An exploratory study of pull-down menu selection as compared to button bars. Technical report CAR-TR-764, University of Maryland, Dept. of Computer Science, Human-Computer Interaction Lab. 1995.

Maddix F.: Human computer interaction: theory and practice. Ellis Horwood, West Sussex 1990.

Norman K.L.: The psychology of menu selection: Designing cognitive control at the human/computer interface. Ablex, 1991.

Sager N., Bross I.D., Story G., Bastedo P., Marsh E., and Shedd D., Automatic encoding of clinical narrative., Comput Biol Med 1982 Vol 12 pp 43-56

Silverberg S.G.: SNOMED encoded surgical pathology databases: 's no big deal – or is it? Modern Pathology 1996 Vol 9 pp 953-954.

Spyns P. and De Moor G.: A Dutch medical language processor. International journal of bio-medical computing 1996 Vol 41 pp 181-205

Wieggers J.G., De Bruijn L.M., Hasman A., and Arends J.W.: De gebruikersinterface voor de invoer van PALGA classificaties. (in Dutch). Accepted for publication in: Medisch Informatica Congres 1997.

APPENDICES

Appendix 2 published as:
De Bruijn L.M., Verheijen E., Hasman A., Van Nes F.L., and Arends J.W.: Speech Interfacing for Diagnosis Reporting Systems: an Overview. Computer Methods and Programs in Biomedicine 1995 Vol 48 pp 151-156

APPENDIX 1: PATHOLOGY DIAGNOSIS AND CODING

1. Categories of diagnostics

Diagnostics are categorised at the department by histology, cytology and autopsy:

- ◊ histology examinations: tissue material (histos = woven tissue, logos = study)
- ◊ cytology examinations: cell material in a fluid environment (cytos = cell)
- ◊ autopsy: post-mortem examination.

A look at the annual report of the pathology department at the Academic Hospital in Maastricht reveals the distribution of examinations with the departments that request the examinations. Table 1 summarizes these figures, and the development from 1990 to 1995. The table shows a steady increase of histology examinations and a constant level of cytology examinations.

TABLE 1: pathology examinations at the Academic Hospital in Maastricht. Source: annual reports 1992 - 1995

	1990	1991	1992	1993	1994	1995	avg	avg growth pa (sd)
TOTAL	23896	23170	25258	26753	27161	26578	25469	536 (1244)
histology	10511	10752	12767	13513	14467	14670	12780	832* (736)
surgery	2237	2311	2468	2299	2498	2444	2376	41 (152)
dermatology	1235	1161	2213	2897	2955	2943	2234	341 (500)
gynaecology	1231	1192	1085	1205	1272	1192	1196	-8 (97)
intern. med.	2019	2307	2698	2731	3110	3498	2727	296** (153)
urology	761	742	815	897	1084	1128	905	73** (75)
other	3028	2994	3488	3502	3548	3465	3338	87 (232)
cytology	13046	12051	12151	12924	12403	11628	12367	-283 (719)
gynaecology	9152	8290	8075	7869	7550	7182	8020	-394** (270)
other	3894	3761	4076	5055	4853	4446	4347	110 (552)
autopsy	339	367	340	316	291	280	322	-12 (23)

*: significantly ≠0, p=.90 **: significantly ≠0, p=.95

Natural language in reporting 2.

The following paragraphs give a global description of the looks and characteristics of free-text reports as they are produced to communicate histology diagnoses. For these descriptions, the collection of reports was analysed that also served in the other experiments in this thesis: a collection of 7500 histology reports.

The collection of reports was retrieved from the archives of two laboratories. These were all histology reports; 2500 from Elkerliek Hospital (the 'ELK' collection) and 5000 from the Academic Hospital in Maastricht (the 'AZM' collection). In both cases, this was the whole production of about the first three months of 1994. All cases were included in the experiments and analyses - no reports were left out. The reports were dictated by 25 pathologists and residents.

This section describes some characteristics of the texts, both describing the average features of a common histology report and exploring the differences between the two originating laboratories. In order to compare the sets, the larger (AZM) set is split in two halves (chronologically - AZM1 and AZM2) so that we now have three sets of equal size.

Report length 2.1

The distribution of report length is very skew, so averages are better given in medians than in means. Table 2 shows the length of reports for the collections, along with the quartile values. Reports from ELK differ significantly in length from AZM reports (two-sample t-test; $t=10.0$). The distribution of these report lengths is plotted in figures 1a and 1b

Vocabulary 2.2

In the total set of reports, about 20,000 different words occur. This includes inflections of verbs, nouns, adverbs and adjectives, variations in spelling preferences, abbreviations, numeral values and also some words with misspellings or typing errors.

TABLE 2: report length (number of words) for ELK and AZM collections

	median *(mean)*	1st quartile	3rd quartile	max. report length
ELK	100 *(126)*	68	149	706
AZM1	127 *(154)*	92	184	1168
AZM2	125 *(151)*	89	175	1318

FIGURE 1A: distribution of report length for collections ELK and AZM1

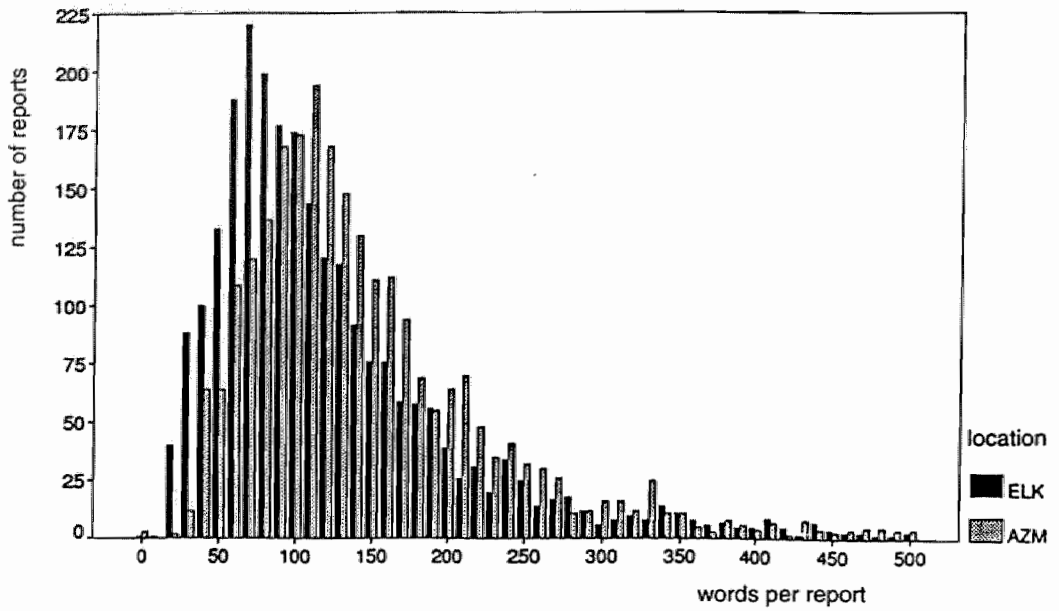
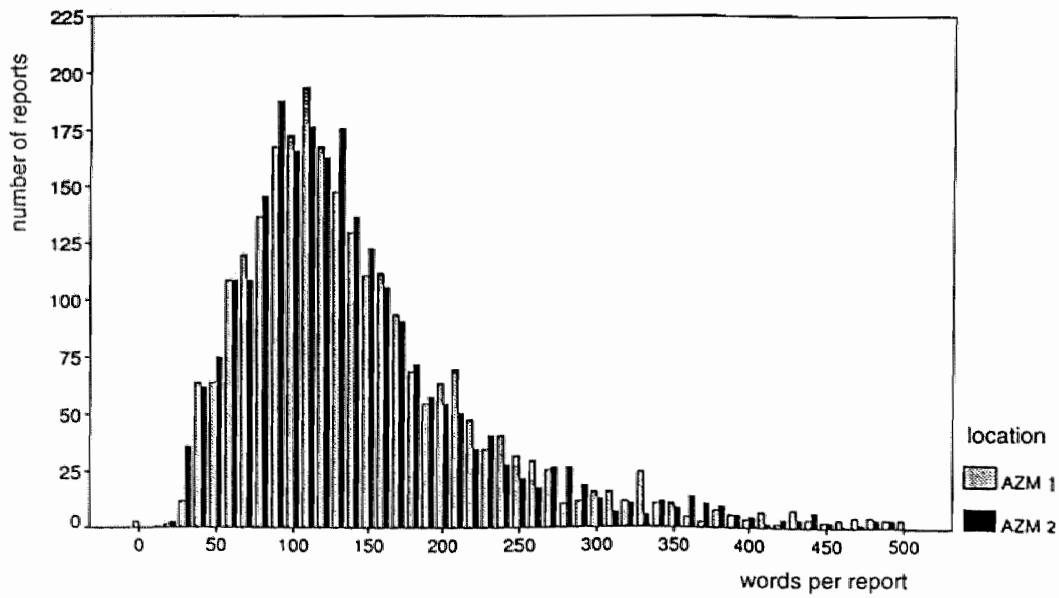


FIGURE 1B: distribution of report length for collections AZM1 and AZM2



The words in the reports follow the general rule that a very small number of different words is responsible for a large number of occurrences.

A number of observations:	a total of about 1.09 million words was encountered; there were about 20,000 different words;
8200	of these occurred only once
2600	occured only twice
15,300	different words occurred 10 times or less
4600	words occurred more than 10 times
1200	words occurred more than 100 times
163	words occurred more than 1000 times
10	words occurred more than 10,000 times (een, de, van, met, en, het, in, is, i, geen)

Figure 2 summarises this data: it shows a Zipf-curve (Zipf 1949); the word frequency put against the rank of the words. At the left end of the curve, the very-frequent words are displayed with their frequencies:

een	40361	met	27667	in	22082	i	11416
de	36485	en	26869	is	13681	geen	11363
van	34231	het	22965				

At the right end of the curve, the 8200 words with f=1 can be seen, next to the 2600 words with f=2, etc.

Figure 3 displays the coverage of occurences if words are inserted one by one. It shows that 20% of the text is covered with only 10 words; with 50 words a little more than 40% is covered and just less than 85% of the text is covered with 1000 words. But 100% is slowly approached on adding up to 20.000 words.

Figure 4 shows how the number of different words increases with the number of reports (which were chronologically inserted). The curve shows clearly that even after the 7500th report, the number of different words does not begin to level out. The observations in figure 4 can be modeled with a $y=a+b*x^c$ curve. With optimum parameter estimations (a=–1840, b=730 and c=0.382), a near perfect fit could be made ($R^2=.99986$).

If this model is extrapolated, then upon entering the 15,000th report, 26,900 different words may be expected, after 50,000 reports there would be about 43,700 different words (100,000 → 57,500; 500,000 → 107,900; 1,000,000 → 141,200).

A model estimated on the first half of the observations (up to 3750 reports), gives a total $R^2=.9989$; at 7500 reports the estimate differs with just over 1% with the observation and the prediction for the 1,000,000th report differs less than 10% with the prediction of the previously mentioned model.

Categories of words 2.3

Function words form the cement in 'sentence-bricklaying' - they connect the content words. Most of them are short words that occur frequently.

FIGURE 2: Zipf curve of words in the pathology reports

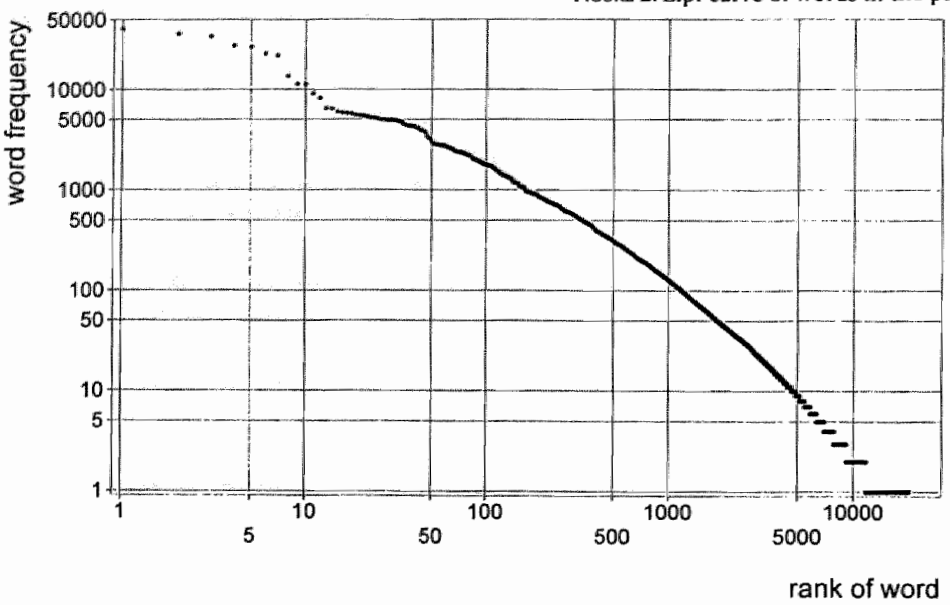
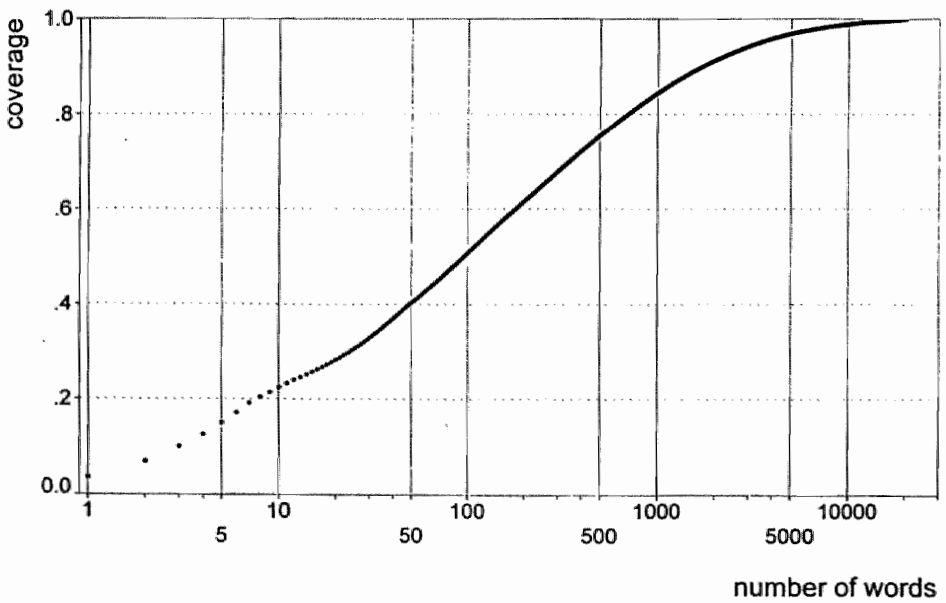


FIGURE 3: Proportion of words in the texts covered with increased number of words (e.g. 50% of all word occurrences is covered with only 100 different words)



On the collection, frequency counts on several function words showed:

category	*different words*	*occurences*	*examples of words*
◊ articles	3	99800	de, het, een (the, a)
◊ pronouns	15	134400	van, met, in, door etc (from, in, with etc)
◊ auxiliary verbs	23	30700	is, zijn, wordt, worden (is, are)
◊ personal pronouns	2	5900	we, ik (we, I)
◊ conjunctions and adverbs	43	105700	en, er, als, ook, etc (and, if, too, etc)

These few words are responsible for more than a third of the total of words.

A comparison between the vocabularies of the AZM reports and the ELK reports reveals that some words are used equally in both laboratories, others are used (almost) exclusively in one location. This can be rated by dividing the word frequencies for the two laboratories (if the word frequency is 0, a value of 1 is inserted to allow the calculation of the ratio. Since only the large differences are assessed, this gives only a minor aberration).

Figures 5a and 5b display the distribution of word frequency ratio. If the word frequencies are about the same, their ratio lies around 1 (which gives a natural logarithm of 0). These words are contained in the middle bars. On comparing the two halves of the AZM collection, one sees that there is some chance that a word occurs up to 20 times more often in one halve than in the other. There are 3 words, that occur even up to 90 times more often in one halve.

The comparison of ELK with AZM1 also shows a pyramid, but with a much broader basis. There are 109 words that are highly typical for the AZM1 collection but are rarely or not used at ELK. The inverse is the case for 74 words. These differences are caused by reporting style of the pathologist, and sometimes by the spelling preference of the secretaries.

FIGURE 4: Increase of corpus vocabulary with addition of reports. New words are encountered even after 7,500 reports

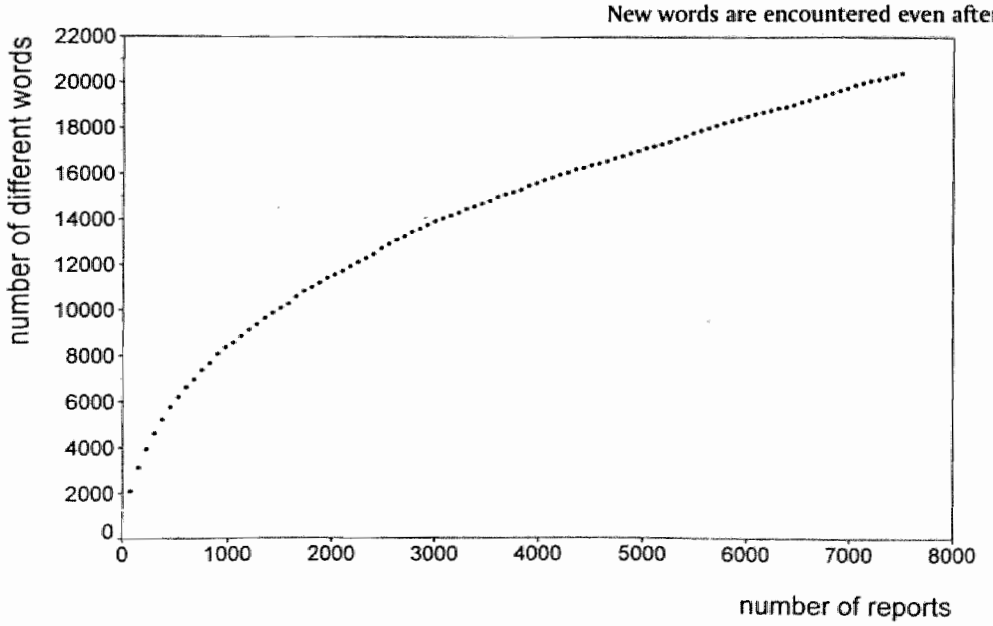


FIGURE 5A: word frequency ratio bar chart. Bars represent number of words (8484) that have (almost) the same word frequency (ln ratio=0) in ELK and AZM1 archives, up to words that occur 400 times more often (ln ratio=6) in ELK archive (3) or in AZM1 archive (8)

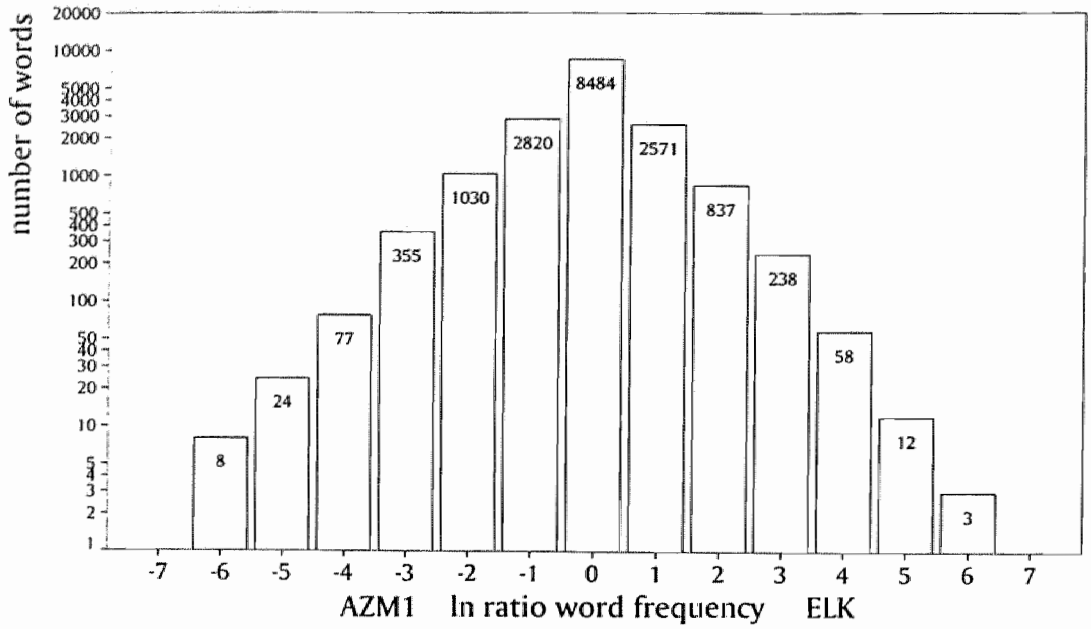
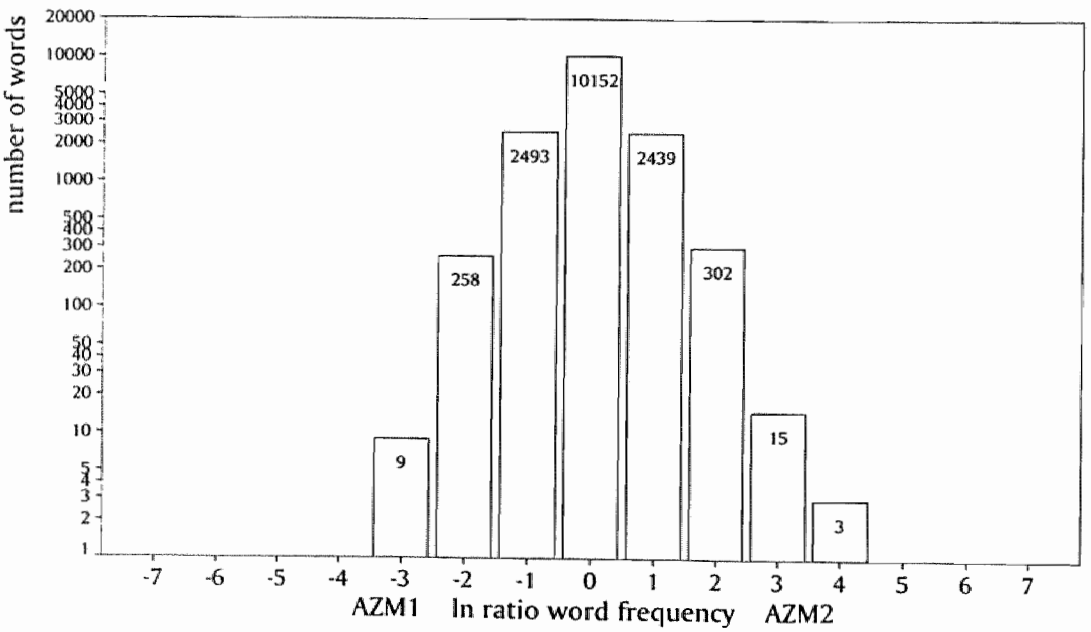


FIGURE 5B: word frequency ratio bar chart for two halves of AZM archive. Most words (10125) occurred almost equally (ln ratio=0) in the two halves



The use of classification codes 3.

Similar analyses were made for the frequencies of occurrence for SNOMED concepts and codes. A SNOMED code is a 6-character string, e.g. P11400. A concept may be a concatenation of strings, e.g. M55720E50940.

In the collection of 7500 excerpts, a total of 1344 different concepts was found in a total of 32195 concepts (on average, 4.29 concepts per excerpt). The ten most frequent concepts were:

1.	P11400	f=2770	biopsy
2.	P11000	2504	resection
3.	T01000	2416	cutis
4.	TYY980	884	left sided
5.	M09450	877	no malignity
6.	TYY990	863	right sided
7.	M00100	784	no pathological change
8.	M43000	766	chronic inflammation
9.	M42000	380	chronic active inflammation
10.	M87500	365	naevus dermalis
	55	>= 100	(55 different concepts occurred more than 100 times)
	112	>= 50	
	371	>=10	
	49	=5	(49 different concepts occurred 5 times)
	77	=4	
	102	=3	
	181	=2	
	440	=1	(440 different concepts occurred once)

In the collection a total of 1237 different codes was found in a total of 34273 codes (on average, 4.57 codes per excerpt). The ten most frequent codes were:

1.	P11400	f=2773	biopsion
2.	T01000	2616	cutis
3.	P11000	2583	resection
4.	TYY980	887	left sided
5.	M09450	877	no malignity
6.	M43000	871	chronic inflammation
7.	TYY990	867	right sided
7.	M00100	784	no pathological change
9.	T84000	667	endometrium
10.	T04000	516	breast
	58	>= 100	(55 different concepts occurred more than 100 times)
	118	>= 50	
	378	>=10	
	42	=5	(49 different concepts occurred 5 times)
	76	=4	
	95	=3	
	169	=2	
	359	=1	(440 different concepts occurred once)

Reference

G.K. Zipf: Human Behaviour and the principle of least effort. Addison Wesley 1949

APPENDIX 2: SPEECH INTERFACING FOR DIAGNOSIS REPORTING SYSTEMS: AN OVERVIEW

1. Reporting methods: advantages and disadvantages

Several methods are commonly used in diagnosis reporting, and others are potentially interesting. Methods can be distinguished by the process of report production, which partly determines the final form of the report. Free text composition and coded language composition are the two most frequently used methods in practical settings. Dictation is preferred in report production because speech – as opposed to typewriting – leaves the eyes free for examining the photo or tissue, and the hands free for using the magnifying glass, for handling tissue, or for adjusting the microscope.

First, reports can be written in *free text composition*. The report is dictated and transcribed into a written version. The advantage of free text composition is that the diagnosis is tailor-made for the case in all its details and nuances. The disadvantage is that composition is rather time-consuming, as is transcription of the spoken report. An alternative to the written final form of the report would be a spoken report that is sent to the physician, or that is accessible by telephone. However, a written report is more easily accessible, and is better suited for automated storage and retrieval (Jost 1986).

Second, reports can be written in *coded language*, or canned language. Coded languages evolved from the use of abbreviations for labelling routine cases in full formal languages that describe diseases in anatomic, etiologic and descriptive code words (cf. Jost 1986, Lehr et al. 1973). Sometimes parameters must be added to the standard description to adapt it to the specific case. After generation, the coded terminology can be expanded into standard sentences in natural language, forming a report to be sent to the physician. The advantages are that with a coded language, reports are generated very quickly, and automation can support the generation of a report. Automation has been introduced in various ways, such as form filling, graphical display input (Jost 1986), bar code entry (Choplin et al. 1984) and speech recognition. The Kurzweil VoiceMed systems are based on coded language report generation. VoiceMed systems have found widespread daily use, especially in radiology laboratories. A faster throughput of reports and cost reduction in the transcription process have been reported by Choplin et al. (1984), Robbins et al. (1987) and Sijtsma & Zweekhorst (1992).

Disadvantages of coded language are, that it requires a relatively long learning phase before the diagnostician is familiar with the various codes. Even then, it works best with standard cases. Robbins et al. (1987) reported a 70% coverage

of radiology reports for Kurzweil Voice Rad, the other 30% must be reported with traditional dictaphone – free text composition. Then, an exceptional case requires an exceptional reporting routine so that the radiologist must put additional effort into both diagnostic reasoning and reporting. This mental load may have a negative effect on the quality of the diagnostic reasoning process. It also entails the danger that a slightly deviant case is treated as its nearest routine case so that its diagnosis can be composed with the system anyway. No research covering this subject has been reported. Furthermore, the automated reporting support often *guides* rather than *follows* the doctor's diagnosis strategy. It is not clear what effect this has on the quality of the diagnoses. Finally, the reduction in composition time that is achieved with a coded language also brings a reduction in redundancy. In other words: a small error in coding or transcription may result in a syntactically legal (but semantically different) report. Such errors are very difficult to detect, and may have serious negative consequences.

Report generation and storage by graphical representation is an option that becomes more interesting with the recent developments in high quality graphical displays. Easy to use interfaces can be applied, such as input by pointing. Although developments in graphical representation are very interesting, they fall somewhat outside the scope of this article.

In pathology, the boundary between routine cases and rare cases is estimated to be less distinct than in radiology. Even routine cases may differ in subtle details. Specific research is in progress to assess this claim. The pathology department of the Academic Hospital Maastricht – one of the partners in this project – has therefore always held on to free composition full text reporting.

2. Applying a perfect speech recogniser

Transcriptions are made by secretaries, who return the typed reports to the diagnostician for signature. This step takes half a day in the logistics of most diagnostic departments. Furthermore, transcription is a task that demands typing skill, knowledge of medical terminology (spelling, pronunciation), and familiarity with the doctor's dictating style. A transcription department is relatively large and expensive, is highly routine based, and vulnerable to illness, leave and turnover of the typists. All this makes the transcription process a subject well worth for innovation.

Automatic speech recognition allows that free-text reporting is maintained, but combines it with the greater speed and lesser cost of automated systems. Furthermore, a diagnosis can be delivered quicker to the physician if turn-around time through the transcription department is saved. Choplin et al. (1984) reported a reduction in transcription staff by 30% to 50% after

introduction of a (bar-code oriented) radiology reporting aid, and a reduction of some 35% in turnaround time. The effect of a quicker diagnosis may be higher, eg. if a patient's treatment can be started or continued sooner. In normal situations, a transcribed report is returned to the diagnostician for correction or verification, often on the day after dictation. If the diagnostician could perform this step immediately after a diagnosis session, the cases would still be fresh in the memory, and in case of doubt the tissue section or photo can be reviewed. Therefore, the contents of a report that is composed quicker may be more dependable. An experiment on this subject is in progress within our project (Verheijen et al. 1997).

An ideal speech recogniser can be defined as fast, unlimited and errorless. Even with such a recogniser, problems can be expected in interaction.

Users are not accustomed to speaking to a machine, and are known to adapt their speaking habits (Isaacs et al. 1993, Kennedy et al. 1988). Talking to a computer may even give them unrealistically high expectations about the capabilities of the system, or psychological problems such as uncertainty in interaction. The application of speech recognition in this setting is not expected to be vulnerable for these drawbacks, because the present situation – speaking to a dictaphone – is highly similar to it.

Feedback requirements for a speech recognising system are not totally clear yet. At least, the system should offer similar feedback mechanisms as dictaphones (rewind, playback and rerecord) to support the task of dictating itself. For error detection and correction during dictation, Shurick et al. (1985) suggest visual feedback rather than auditive feedback when the task is performed under time pressure; if data errors can be captured later in the process, *no* feedback may even be appropriate. Visual feedback during the task can be accomplished by, eg., projecting the transcribed text into the ocular field of the microscope, but its desirability can be questioned. Specific research is to be performed on this subject. Other feedback would support system interaction, but is too dependent on actual system design to be adequately discussed here.

If a diagnostician has access to a speech-controlled transcribing device, it may be embedded into a total speech-accessible diagnostic support system. To put it even stronger: a doctor may be reluctant to accept the extra tasks in daily work, that are inherent with the introduction of a speech recogniser, if there are no additional features like diagnostic support (Sijtsma & Zweekhorst, 1992). Speech as an interaction mode has the advantages that it is a natural communication method with a relatively high bandwidth (Martin 1989, Damper & Leedham 1992). Then, diagnosticians – often inexperienced with computer use (cf. Shiffman et al. 1991) – have access to data communication, information

retrieval and expert system facilities. These facilities, combined with the accessability of nationwide central archives and the availability of demographic data, can support the doctor with relevant information in solving difficult cases. In the present situation, doctors are often restrained to request additional information because they are aware of the costs in hassle and time of such requests (personal communications).

Note that interaction characteristics differ between full text dictation and interactive retrieval of information with short commands. Recognition demands, especially concerning speed and accuracy of recognition, will change with the interaction mode (cf. Martin 1989). An important issue here is, that recognition of dictated text is analogous to the present situation, whereas speech interaction with a system is a new application. The latter involves the rather odd activity of addressing yourself directly to a machine (cf. Shiffman et al. 1991). Interface requirements may not be generalised between these two tasks.

Applying a realistic (imperfect) speech recogniser 3.

State of the art technology does not supply speech recognisers that approach ideality. Limitations occur in processing speed, vocabulary capacity and recognition rate. However, even with these limitations, automatic speech recognition and diagnostic reporting could make a harmonious combination.

Speech recognition systems do not make spelling errors. Other errors *do* occur, resulting in typical error rates between 2% and 10%. The following error categories can be distinguished:

- ◊ an utterance was not detected and thus not recognised.
- ◊ an utterance *was* detected but not recognised: no match was found in the vocabulary.
- ◊ a word was recognised while nothing was said. This may occur when the microphone picks up background noises or when the speaker coughs or says 'uhm'.
- ◊ a spoken word is incorrectly recognised. This sometimes occurs when words sound alike (e.g. horse and norse). A more difficult problem is dealing with homophones – words that sound exactly the same, e.g. horse and hoarse. Good vocabulary design and/or contextual correction algorithms can reduce these substitution errors.

Of course, speech recognising devices should be subject to high quality standards when applied in medical information systems. Therefore, error rates must be brought down towards 0% by comprising features to support recognition error correction, either automatically or by hand. A confidence figure can be calculated for each recognised utterance. If that confidence figure falls below a

certain threshold, the suggested recognised text can be marked for doubtfulness, e.g. by underlining or colour coding the displayed text. Editing support can be provided to the user, such as easy selection of doubtful phrases, offering alternative suggestions and offering playback of the speech signal. Note that the editing work may be done by a secretary: speech recognition then still saves much work in transcription. The Philips dictation system is designed to be used by a transcriber and not by the speaker itself (Raaijmakers 1993).

Without going deep into the technical implementation of speech recognition algorithms, a few words on the topic are well spent here. Basically, there are two different approaches to speech recognition: template based recognition and phoneme based (Markov chain based) recognition (Helander et al 1988). In template based recognition, a prototype speech representation is available for each word in the vocabulary. An utterance is matched against the library of prototypes, and the most likely match is accepted as the text equivalent of the utterance. In phoneme based recognition, a speech signal is divided into short time chunks (typically about 20 ms each). Per time chunk, the most likely phoneme is selected from a set of 25 to 35 basic phonemes. A series of phonemes can then be transformed into a word or a string of words. See e.g. Helander et al. (1988), Lee (1989) or Reddy (1990) for an overview of technical aspects of speech recognisers.

The choice of a recognition method is related to some user aspects of the system. First, a template based system requires that words are separated with silences. In phoneme based systems, recognition is facilitated with isolated speech, but connected speech recognition is also possible. Gould et al. (1983) concluded from experiments that for ease of use "Isolated word speech with large vocabularies may be nearly as good as connected speech systems for a listening typewriter". It does, however, demand habituation by the speaker.

Second, a template based system demands explicit training of each word in the vocabulary. Some systems are shipped with a preprogrammed word/prototype database. Possibly, generalisation algorithms can substitute specific training of new words. Phoneme based recognition may have inherent generalisation for new words.

Third, many template based systems are speaker dependent, or at best speaker adaptive. That means that new users must perform some initiation sessions before the system can recognise their voices. A Markov chain recogniser must at the most adapt to a new user's phoneme set, which is often accomplished by the user reading a few lines of training text.

Fourth, template based systems seem to be less error prone than phoneme based systems. This is, however, also strongly dependent on vocabulary size, vocabulary design, speaker characteristics and individual word properties. Correction algorithms can reduce the error rate significantly.

Fifth, the performance of a template based system is dependent upon the size of the vocabulary, which is less so in phoneme based systems. In a template based system, a recorded utterance is matched against each template in the vocabulary. Speed and accuracy of this process may decrease with larger vocabularies. However, performance may be restored by carefully designing a set of smaller sub-vocabularies.

In conclusion, the choice between a template based system and a phoneme based system can only be made after estimation of the user's demands, the range of the system's application, the preparedness to customise the system, and the costs.

Despite the difficulties in speech recognition technology, implementation in diagnostic departments seems to be advantageous:

- ◊ the diagnostician's office offers as good acoustical conditions as one may desire: a noiseless environment, an immobile speaker and good possibilities for microphone placement.
- ◊ the speaker is used to dictating, so stuttering, coughs and hesitations may be relatively few, or may be deleted by the user with sound editing facilities comparable to a dictaphone's (e.g. rewind, replay and rerecord).
- ◊ the set of different speakers is small, and remains constant over a longer period of time. A system may profit from the benefits of speaker dependent or speaker adaptive algorithms.
- ◊ real-time recognition is not necessary. The system should be faster than the processing time of the average transcription department. A pathologist is satisfied when the first report of a dictating session is already processed when the whole session is finished (this would mean some 5x real-time recognition). Non-real-time recognition would allow algorithms other than 'from left to right' to obtain a higher recognition rate.
- ◊ vocabulary and syntax in diagnosis reporting seem to be strictly defined, albeit not formally modelled. A large archive of reports is usually available, so realistic models for different categories of reports can be derived from it. This would enable multi-level recognition, in which acoustic features of an utterance are used to generating suggestions, that are then matched against the probabilities of words occurrence, given the context (Rabiner & Levinson, 1990). Further study on this subject is being performed in our project.

4. Concluding remarks

Speech recognition as a whole is promising in diagnosis reporting support. Various techniques differ in applicability, so a balance must be found between user needs, modes of report generation, error handling, system complexity and

cost. The preferences in these will differ from laboratory to laboratory. The high-end of speech recognition – a fast, continuous speech recogniser with no limits on vocabulary or syntax – is not available as it is, but advances in technology seem promising enough to start preparing interface definitions of such a system.

References

Choplin R.H., Boehme J.M., Cowan R.J. et al.: A computer-assisted radiologic reporting system. Radiology 1984 Vol 150 pp 345-348.

Damper B. and Leedham G.: Human Factors. In: Rowden C: Speech Processing. McGraw-Hill 1992

Gould J.D., Conti J., and Hovanyecz T.: Composing letters with a simulated listening typewriter. C-ACM 1983 Vol 26 pp 295-308

Helander M., Moody T.S., and Joost M.G.: Systems design for automated speech recognition. In: Helander M.: Handbook of human cumpter interaction. Elsevier North Holland 1988 pp 301-319

Isaacs E., Wulfman C.E., Rohn J.A. et al.: Graphical access to medical expert systems: IV. Experiments to determine the role of spoken input. Methods of Information in Medicine 1993 Vol 32 pp 18-32.

Jost R.G.: Radiology Reporting. Radiology Clinics of North America 1986 Vol 24 pp 19-26

Kennedy A., Wilkes A., Elder L. et al.: Dialogue with machines. Cognition 1988 Vol 30 pp. 32-72.

Lee K.F.: Automatic Speech Recognition: the development of the SPHINX system. Kluwer 1989.

Lehr J.L., Lodwick G.S., Nicholson B.F. et al.: Experience with MARS (Missouri Automated Radiology System). Radiology 1973 Vol 106 pp 289-294

Martin G.L.: The utility of speech input in user-computer interfaces. Int. J. Man-Machine Studies 1989 Vol 30 pp 355-375

Raaijmakers R.: Philips systeem noteert dictee. Polytechnisch weekblad 1993, issue 36.

Rabiner L.R. and Levinson S.E.: Isolated and Connected Word Recognition - Theory and Selected Applications. In: Waibel A & Lee KF: Readings in Speech Recognition. Morgan Kaufmann 1990 pp 267-296

Reddy D.R.: Speech recognition by machine: a review. In: Waibel A. and Lee K.F.: Readings in Speech Recognition. Morgan Kaufmann 1990.

Robbins A.H., Horowitz D.M. et al.: Speech-controlled generation of radiology reports. Radiology 1987 Vol 164 pp 569-573

Shiffman S., Wu A.W., Poon A.D. et al.: Building a speech interface to a medical diagnosic system. IEEE Expert 1991, Vol 6 pp 41-49

Shurick J.M., Williges B.H. and Maynard J.F.: User feedback requirements with automatic speech recognition. Ergonomics 1985 Vol 28 pp 1543-1555.

Simon M., Leeming B.W., Bleich H.L. et al.: Computerized radiology reporting using coded language. Radiology 1974 Vol 113 pp 343-349

Sijtsma W., Zweekhorst O.: Using speech when creating medical reports: Kurzweil's VoiceReport/VoiceMed. ITK/Think quarterly 1992 Vol 1 pp 64-66.

Verheijen E., Van Nes F.L., De Bruijn L.M. et al.: Automatic Speech Recognition in medical environments: will it improve report quality? Accepted for publication in Behaviour and Information Technology, 1997.

Summary · Samenvatting

SUMMARY

Tissue that is sampled or excised at an operation, is sent to the pathology department, where it is microscopically examined. A diagnosis on illness is made, or the presence of illness is ruled out. As such, pathology is one of the diagnostic departments (some will say: *the* diagnostic department) in a hospital. For making the diagnosis, pathology is supported by the scientific discipline that bears the same name.

The result of a pathology examination, either for scientific or clinical reasons, is knowledge. In both environments, knowledge is meaningless without it being recorded and reported. The patient's physician gains important information from the pathologist's report so that the treatment of the illness can start or continue. The report may later help the pathologist to determine the course of a process or to assess new conditions in the same patient, and to interpret similar patterns when these are observed in another patient. Information from pathology examinations helps other scientists to do their studies.

Pathology examinations are reported in a piece of text that describes the observations and the conclusions. Despite the medical 'mumbo-jumbo' in it, it is said to be in *natural language*. The international medical world is not always helped with this free format, so a system of well-defined terms was developed: SNOMED (Systematized Nomenclature of Medicine). Any diagnosis can be described in it. Dictionaries are available to translate the SNOMED terms from one language to another, which makes data internationally exchangeable.

Natural language and SNOMED are used in parallel: the natural language report is clearer and easier for the human reader, while the SNOMED classification of the same case is very well suited to store and retrieve cases in a computer database, and to select groups of cases in the database.

It is difficult and burdensome for a pathologist to write a good SNOMED classification. The classification language is vast: the Dutch version has about 16,000 terms. Errors occur – this has been shown in literature (see chapter 2) – and even if they rarely occur, they decrease the value of the total database. Different pathologists may classify a single case with different SNOMED terms, depending on their personal preferences. This endangers the idea of standardization.

Automatic classification has been coined to solve these problems. In Chapter 2, studies are presented that attacked automatic classification by using language models to analyze text, or by using medical knowledge models to interpret reports, or by using dictionary-lookup of medical terms. These studies have not led to widespread use for various reasons.

The answer to this thesis' question – *'is it possible to automatically assign classification codes to a diagnosis report that is written in natural language'* – was sought in a different approach in automatic text classification. This method considers a new report against a large collection of other reports. If the collection contains another report that looks very similar to the new report, the classification of the older report may also apply to the new report. Two conditions should be met:

1. you need a collection of reports that are already classified. This is no problem: most laboratories have used computer systems for a long time;
2. you need a measure of similarity (between two texts) that can predict the true medical similarity between two cases.

Most of this thesis concentrates on the second condition: the definition of text similarity, and evaluation. This approach of searching a collection to classify new cases, is called the Nearest Neighbor method. It is born from methods in Pattern Recognition and Information Retrieval. The appeal of the method is the separation between method and domain: you can classify texts from another discipline or language by using another collection, and if the collection is kept up-to-date, gradual changes in reporting style or spelling preferences are automatically taken up. The Nearest Neighbor method does not distinguish between difficult cases and easy cases, although it must be said that is easier to find very similar reports for common cases than for rarer cases.

The text similarity is calculated by comparing the words that are used in one report with the words in the other report. Words that are shared by both reports, *raise* the similarity measure, and words that occur in only one of the reports *lower* the similarity measure. The result is normalized, so that exactly identical texts have a similarity = 1, and entirely different texts have similarity = 0, regardless of the length of the texts. Words should not contribute equally to the similarity measure: a rare medical term that occurs in two reports says more about the true medical similarity than words such as 'the', 'is' or 'to'. Weight schemes balance these differences. Chapter 3 presents the details of the Nearest Neighbor method and gives the result of a pilot experiment.

Words that *mean* the same may *look* different, so that they lower the text similarity instead of raising it. Dutch language is notorious for 'pasting' words together (e.g. 'vreemdlichaamreuscelcarcinoompje'). The syntactic context may alter the looks of a word (e.g. 'grijs' and 'grijze'), as do spelling preferences (*ph/f*, *ae/e*, *c/k*). It was found that performance does increase when words are transformed into their base forms, but this improvement was so slight that it is hardly worth the trouble (see chapter 3).

Other variations in the model and its application are treated in chapter 4. It was found that different laboratories may have different reporting styles: new reports often homed for nearest neighbors that came from the same

'mother'-laboratory even when the collection was mixed. If the nearest neighbor was forced to come from the other laboratory, performance was reduced. It was also found that a comparison with the whole report gave better results than comparison with only the 'conclusions' section report. Performance gets better with a larger archive, but even the collection of 7,500 reports did not cover the wide range of pathology. Finally, the initial choice of using *words* to form the basis of the similarity computation was tested against an alternative of using letter *n*-grams (sequences of *n* letters that occur in both texts). Words proved to be better suited than *n*-grams.

In the previous paragraphs, *performance* was referred to without explaining how evaluation takes place. In evaluating the experiments, the judgment of experts was assumed to be the golden standard. A large scale experiment was run in which a group of 18 pathologists rated whether a suggested SNOMED classification line for a given case report was well suited, less well suited or poorly suited. These results indicated that for just under 50% of the reports, the method worked succesfully (see chapter 5). Based on a number of arguments, it is defended in this chapter that a success percentage of 66% gives a more realistic figure. The expert evaluation pointed out a number of weak spots in the method, notably concerning reports with multiple diagnoses, reports that described skin tissue and reports that contained specified key terms.

Experts are not always available for evaluation. Therefore an alternative evaluation method was developed, which is considered to be a *silver standard* rather than a golden one. Diagnosis reports from the laboratory, which are therefore supplied with a SNOMED classification line (written by the pathologist on duty), can be used as test reports in simulations. Before the simulation, the classification line of a test report is separated from the text. Suggestions for classifications are sought on the basis of the text, and these suggestions are then evaluated by comparing them with the original classification of the test report. This Silver Standard was validated on the basis of the expert ratings of the large-scale experiment (chapter 5). This Silver Standard also makes it possible to check whether *good* classifications were left behind in the collection when the Nearest Neighbor method delivers only *poor* classifications – a situation in which the method clearly fails. These situations should be separated from the cases where the collection contains nothing but poor classifications.

The silver standard was put to work in a large scale simulation (6,640 reports). On the basis of earlier observations, a number of enhancements was developed and evaluated for the method, notably to deal with 'skin'-reports, and reports with key-terms ('left'/'right' etc). Chapter 6 describes the outcomes: for about 73% of the cases the collection contained a good classification. It could be

found by the nearest neighbor method within the first five suggestions in about 70% of these cases. For about 96% of the cases, the collection contained a classification that was good or partly good; the nearest neighbor method was able to include it in the first five suggestions in about 80% of the cases.

With these figures, the boundaries of the basic nearest neighbor method have been determined. Improvement may be possible with the adoption of other weighting schemes and searching by clusters. Absolute performance should increase with a larger collection of reports. The figures that were found in the experiments indicate that the nearest neighbor method for text classification is an good competitor for syntactical based, semantical based or index-based classification methods as described in literature (see chapter 2).

The question that was posed at the beginning of this thesis can be cautiously confirmed. Automatic classification is possible but the gap with perfect performance (also considering the importance of perfect performance) is still too high to consider autonomous automatic classification. The nearest neighbor method did show potential to support a pathologist in writing classifications. The computer-assisted pathologist probably gives the best possibility for feeding the national archive with good classifications. Classifications that are correct, complete, consistent and concise.

SAMENVATTING

Weefsel dat bij onderzoek of een operatie wordt uitgenomen, wordt naar de pathologie-afdeling gestuurd voor microscopische beoordeling. Daar wordt een diagnose gesteld op de aanwezigheid of afwezigheid van aandoeningen. De pathologie-afdeling is daarom een van de diagnostische afdelingen (of volgens sommigen *de* diagnostische afdeling) in een ziekenhuis. De basis voor het stellen van diagnostiek is de wetenschappelijke discipline die eveneens Pathologie heet.

Het resultaat van een pathologische beoordeling, zij het voor wetenschappelijke of clinische doeleinden, is kennis. Kennis is zonder waarde als rapportage en opslag ervan achterwege blijft. De huisarts van de patient haalt waardevolle informatie uit het verslag van de patholoog, waarna de behandeling van de patient kan aanvangen of doorgaan. In een latere fase kan het rapport houvast bieden bij het vaststellen van de loop van veranderingen bij dezelfde patient of het verklaren van nieuwe verschijnselen; ook kan het helpen bij het beoordelen van een patient met vergelijkbare beelden. Pathologie-onderzoek verschaft wetenschappers uit andere (medische) disciplines broodnodige informatie voor hun onderzoek.

Pathologie onderzoek wordt gerapporteerd in een stuk tekst waarin de waarnemingen en de conclusies beschreven zijn. Ondanks de medische vaktaal waarmee de tekst doorspekt is, gaat het onder de noemer 'natuurlijke taal'. De internationale medische wereld heeft vaak niets aan dit vrije formaat; daarvoor werd een systeem van nauw omschreven termen ontwikkeld: SNOMED (Systematized Nomenclature of Medicine). Elke diagnose kan erin beschreven worden. In beschikbare woordenboeken is de vertaling naar andere talen van SNOMED termen voorzien, waardoor gegevens eenvoudig internationaal uitwisselbaar worden.

Natuurlijke taal en SNOMED worden gezamenlijk gebruikt: voor de menselijke lezer is natuurlijke taal helder en gemakkelijk, terwijl voor het opslaan van diagnoses in een computer-systeem en het terughalen ervan de SNOMED classificatie meer mogelijkheden biedt.

Het is moeilijk en belastend voor de patholoog om een goede SNOMED classificatie te schrijven bij een geval. De classificatietaal is zeer uitgebreid: de Nederlandse versie van SNOMED bevat circa 16.000 termen. Bij het classificeren worden fouten gemaakt – dit wordt in de literatuur onderkend (zie hoofdstuk 2) – en zelfs als fouten zeldzaam zijn, verminderen ze de waarde van de hele database. Wanneer verschillende pathologen eenzelfde geval classificeren, dan is er een gerede kans dat, afhankelijk van persoonlijke voorkeuren, verschillende SNOMED termen gebruikt worden. Dit brengt het streven naar standaardisatie in gevaar.

Als oplossing voor de genoemde problemen is automatische classificatie aangedragen. In hoofdstuk 2 wordt een aantal onderzoeken aangehaald waarin automatisch classificeren werd aangepakt middels taalkundige analyse van teksten, of het interpreteren van rapporten via medische modellen, of het opzoeken van medische trefwoorden. Om verschillende redenen hebben deze studies niet geleid tot praktisch goed toepasbare systemen voor automatische classificatie.

Voor de onderzoeksvraag die centraal staat in dit proefschrift – *is het mogelijk om automatisch classificatie-codes toe te kennen aan rapporten die in natuurlijke taal opgesteld zijn* – werd het antwoord gezocht in een andere benadering. In deze methode wordt een nieuw rapport beschouwd tegen de achtergrond van een grote verzameling andere rapporten. Als de verzameling een ander rapport bevat dat sterk lijkt op het nieuwe rapport, dan kan de classificatie van het oudere rapport van toepassing zijn op het nieuwe rapport. Daarbij dient aan twee voorwaarden voldaan te zijn:

1. je moet kunnen beschikken over een verzameling voorgeclassificeerde rapporten. Dit is geen probleem: de meeste laboratoria slaan al sinds jaren rapporten op in hun computersysteem;
2. je moet een maat hebben waarmee je de gelijkenis tussen teksten meet zodanig dat dit de echte, medische, vergelijkbaarheid tussen gevallen voorspelt.

Het grootste gedeelte van dit proefschrift draait om de tweede voorwaarde: het definiëren van een gelijkenismaat en het evalueren ervan. De benadering waarbij een verzameling van oude teksten gebruikt wordt om nieuwe rapporten te classificeren, wordt 'Nearest Neighbor Method' genoemd. Het stamt van methoden uit Patroonherkenning en 'Information Retrieval'. Een aantrekkelijke kant van de methode is de scheiding die bestaat tussen de methode en het toepassingsdomein: door andere tekstverzamelingen te nemen, kunnen teksten uit andere disciplines of andere talen geclassificeerd worden, en als de tekstverzameling ondergehouden wordt, worden geleidelijke veranderingen in de rapportagestijl en spellingsvoorkeuren vanzelf opgepikt. De Nearest Neighbor methode maakt geen onderscheid tussen makkelijke en moeilijke gevallen, hoewel vermeld moet worden dat het eenvoudiger is om sterk gelijkende rapporten te vinden voor routinegevallen dan voor zeldzamere gevallen.

De gelijkenis tussen twee teksten wordt berekend door de woorden die in de ene tekst gebruikt worden te vergelijken met de woorden in de andere tekst. Woorden die in beide teksten voorkomen, verhogen de gelijkenismaat, woorden die uitsluitend in een van de teksten voorkomen, verlagen de maat. Het resultaat wordt genormaliseerd zodanig dat identieke teksten een gelijkenis=1 hebben, en compleet verschillende teksten een gelijkenis=0, onafhankelijk van de lengte van de teksten. Woorden dienen niet allemaal evenveel bij te dragen aan de gelijkenismaat: een zeldzame medische term die

in beide rapporten voorkomt, zegt meer over de echte, medische overeenkomst dan woorden als 'de', 'is' of 'met'. Daarvoor krijgen woorden een weegfactor. Hoofdstuk 3 beschrijft details over de Nearest Neighbor methode en geeft de resultaten van een voorstudie.

Woorden die hetzelfde betekenen, kunnen er verschillend uitzien zodat de tekstgelijkenis verlaagd in plaats van verhoogd wordt. Dit kan worden veroorzaakt door het aan elkaar plakken van woorden (bijvoorbeeld 'vreemdlichaamreuscelcarcinoompje'), waarin de Nederlandse taal uitblinkt. Ook worden woorden veranderd door de grammaticale context (bijvoorbeeld 'grijs' en 'grijze') en door spellingsvoorkeuren (*ph/f, ae/e, c/k*). De prestatie van classificeren vertoonde enige verbetering wanneer teksten via de basisvorm van de woorden vergeleken werden, maar deze verbetering was dermate klein dat het niet opweegt tegen de moeite die het kost (zie hoofdstuk 3).

Andere varianten van het model en toepassing ervan worden behandeld in hoofdstuk 4. Verschillende laboratoria vertoonden verschil in rapportagestijl: bij nieuwe rapporten werden archiefrapporten gevonden uit het eigen laboratorium hoewel de verzameling ook rapporten van elders bevatte. Als het zoeken geforceerd gebeurde op rapporten van een ander laboratorium, dan verminderde de prestatie. Betere resultaten werden gevonden bij zoeken met hele teksten ten opzichte van het zoeken met alleen de 'conclusie'-paragraaf van rapporten. Bij het uitbreiden van de archiefverzameling verbetert de prestatie, waarbij zelfs een collectie van 7.500 rapporten het domein Pathologie niet geheel dekt. Een verdere test richtte zich op de basis van de tekstvergelijkingsmaat: de op *woorden* gebaseerde maat werd vergeleken met een maat die gebaseerd was op *n-grammen* (tekstdelen van *n* letters die in beide teksten voorkomen). Woorden bleken beter geschikt dan *n*-grammen.

In de vorige paragrafen werd gesproken over *prestaties* zonder dat nog verduidelijkt werd hoe evaluatie plaats vond. Bij evaluatie van de experimenten werd het oordeel van experts gezien als een 'gouden standaard'. In een grootschalig experiment beoordeelde een groep van 18 pathologen of een aangereikte SNOMED classificatie-regel goed, minder goed of slecht van toepassing was bij het erbijgegeven rapport. Deze resultaten wezen aanvankelijk op een suksesvolle werking van de classificatiemethode in net minder dan 50% van de rapporten (zie hoofdstuk 5). Naar aanleiding van een aantal argumenten werd beredeneerd dat een suksespercentage van 66% een realistischer beeld geeft. Via de experimentresultaten werden enkele zwakke plekken in de methode blootgelegd, met name waar het rapporten betreft met meervoudige diagnoses, rapporten over huidpreparaten en rapporten waarin specifieke sleutelwoorden een grote rol spelen.

Experts zijn niet altijd beschikbaar om experimentgevallen te beoordelen. Daarom werd een alternatieve evaluatiemethode ontwikkeld, die beschouwd kan

worden als een *'Zilveren* Standaard'. In simulatie-experimenten kan gebruik gemaakt worden van diagnostiekrapporten uit een laboratorium, die destijds door de dienstdoende patholoog van een classificatie voorzien zijn. Voordat het rapport in de simulatie gebruikt wordt, wordt deze classificatieregel gescheiden van de rest van het rapport en 'apart gelegd' om later vergeleken te worden met de *aangereikte* classificatieregels. De Zilveren Standaard werd gevalideerd op basis van de expert-oordelen uit het grootschalige experiment (hoofdstuk 5). Met de Zilveren Standaard is het eveneens mogelijk om na te gaan of er *goede* suggesties achter gebleven zijn in de collectie in die gevallen dat de Nearest Neighbor methode lauter matige of slechte classificatieregels aanreikte (en waar de methode dus faalde). Wanneer voor een gegeven geval alleen maar slechte suggesties aanwezig zijn in de collectie, dan treft de zoekmethode geen blaam.

In een grootschalig simulatie-experiment (met 6.640 rapporten) werd de zilveren standaard ingezet. Op basis van eerdere waannemingen werd de methode aangevuld met een aantal uitbreidingen, met name om met 'huid'-rapporten en sleutelwoorden rekening te houden. De tests op de methode – met en zonder uitbreidingen – worden in hoofdstuk 6 beschreven: in ongeveer 73% van de gevallen bevatte de verzameling een goede classificatie. In ongeveer 70% van deze gevallen was de Nearest Neighbor methode in staat om ze te plaatsen binnen de eerste vijf suggesties. In ongeveer 96% van de gevallen bevatte de verzameling goede of deels goede classificaties; die werden door de Nearest Neighbor methode bij de eerste vijf suggesties gezet in 80% van de gevallen.

Met bovenstaande getallen zijn de grenzen van de Nearest Neighbor methode in kaart gebracht. Verdere verbetering zou mogelijk kunnen zijn door andere weegfactor-berekeningen te gebruiken en door bij het zoeken een clusterstruktuur te hanteren. De absolute prestatie kan verhoogd worden door een grotere verzameling van archiefrapporten ter beschikking te hebben. De getallen uit de experimenten duiden erop dat de Nearest Neighbor methode kan concurreren met grammaticaal gebaseerde systemen, semantisch georienteerde systemen of van classificatie-methoden die werken met het zoeken van index-woorden (zie de literatuurbeschrijving – hoofdstuk 2).

De vraag die aan het begin van dit proefschrift gesteld werd, kan bevestigd worden, zij het met de nodige voorzichtigheid. Automatische classificatie is mogelijk, maar de kloof naar perfect presteren is nog te groot (zeker gezien het belang van perfect classificeren) om autonome automatische classificatie te overwegen. De Nearest Neighbor methode bewees zich wel in staat om een patholoog zinvolle ondersteuning te bieden bij het opstellen van classificatieregels. Een door de computer ondersteunde patholoog geeft waarschijnlijk de beste garantie om het landelijk archief te voorzien van goede classificaties. Classificaties die correct, compleet, consistent en bondig zijn.

Curriculum Vitæ

1966	-	geboren op 4 december te Eindhoven
1979 - 1985	-	Middelbare schoolopleiding: Atheneum B aan het Eindhovens Protestants Lyceum
1985 - 1987	-	Propaedeuse Electrotechniek aan de Technische Universiteit Eindhoven
1987 - 1991	-	Doctoraalstudie Techniek en Maatschappij aan de Technische Universiteit Eindhoven, specialisatie in mens-computer-interactie bij Techniek en Communicatie, afgestudeerd op een onderzoek naar gebruikersaspecten bij database-interactie.
1991 - 1992	-	dienstplicht: gedetacheerd bij TNO - Instituut voor Zintuigfysiologie (tegenwoordig Technische Menskunde) te Soesterberg, betrokken bij onderzoek naar gebruik van trainingssimulatoren en ontwikkeling daarvan.
1993 - 1997	-	Werkzaam als Assistent in Opleiding bij de vakgroep Medische Informatica aan de Universiteit Maastricht, onderzoek naar rapportage van pathologie-diagnostiek.

PUBLICATIES

De Bruijn L.M., Verheijen E., Hasman A., Van Nes F.L., and Arends J.W.: Speech Interfacing for Diagnosis Reporting Systems: an Overview. In: P. Barahona, M. Veloso and J. Bryant: Medical Informatics in Europe, conference proceedings, Lisbon 1994 pp 546-550.
Reprinted in: Computer Methods and Programs in Biomedicine 1995 Vol 48 pp 151-156.

De Bruijn L.M., Verheijen E., Van Nes F.L., and Arends J.W.: Assigning SNOMED codes to natural language pathology reports. In: J. Brender et al.: Medical Informatics Europe, Copenhagen 1996 pp 198-202.

De Bruijn L.M., Hasman A., and Arends J.W.: Automatic SNOMED classification — a corpus-based method. Computer Methods and Programs in Biomedicine 1997 Vol 54 pp 115-122

De Bruijn L.M., Hasman A., Arends J.W.: Classification of diagnoses that are described in natural language. To appear in International Journal of Technology Management, 1997.

De Bruijn L.M., Hasman A., and Arends J.W.: Automatic coding of diagnostic reports. To appear in: Methods of Information in Medicine.

Verheijen E.J.A., Van Nes F.L., De Bruijn L.M., Hasman A., and Arends J.W.: Interference of Automatic Speech Recognition during diagnostic tasks. In: European Conference on Speech Communication and Technology (Eurospeech '95), Madrid 1995 pp 1279-1282.

Verheijen E.J.A., De Bruijn L.M., Van Nes F.L., Hasman A., and Arends J.W.: Automatic Speech Recognition in the medical environment: will it improve report reliability? To appear in: Behaviour and Information Technology.

Wiegers J.G., De Bruijn L.M., Hasman A., and Arends J.W.: De gebruikers-interface voor de invoer van PALGA-classificaties. (in Dutch). To appear in: Medisch Informatica Congres proceedings 1997.

TENSLOTTE...

Aan collega's, familie, vrienden en anderen – iedereen die mij heeft mee-geholpen in de afgelopen tijd met ideeën, daden, kritiek, begeleiding, afleiding en zelfs verleiding:

mijn hartelijke dank.